Department of Electrical and Computer Engineering
North South University

Senior Design Project

Transformers and BiLSTM-Based Ensemble Modeling for Entity and Relation Aware Biomedical Extractive Question Answering

Md Mehedi Hasan	**2022107 642**
Ahmed Ajmine Nehal	**2013090 042**
Farhana Elias	**2013820 042**

Faculty Advisor:

Dr. Mohammad Rashedur Rahman
Professor
Department of Electrical and Computer Engineering
North South University, Dhaka, Bangladesh.

Summer, 2024

LETTER OF TRANSMITTAL

September 2024

To

Dr. Rajesh Palit

Chairman,

Department of Electrical and Computer Engineering

North South University, Dhaka

Subject: Submission of Capstone Project Report on “Transformers and BiLSTM-Based Ensemble Modeling for Entity and Relation Aware Biomedical Extractive Question Answering”

Dear Sir,

Respectfully, as part of our BSc program, we would like to submit our capstone project report on "Transformers and BiLSTM-Based Ensemble Modeling for Entity and Relation Aware Biomedical Extractive Question Answering". The report uses Transformers with BiLSTM models to generate questions and their answers based on a given context. We gained a lot from this project because it allowed us to gain practical knowledge and apply it in real-world scenarios. To meet all the requirements for this report, we tried our hardest to be as competent as possible.

Please examine this report and provide your valuable input. We would be delighted if you could grasp the problem after reading this report and find it informative and helpful.

We will be highly obliged if you kindly receive this report and provide your valuable judgment. It would be our immense pleasure if you find this report useful and informative to have an apparent perspective on the issue.

Sincerely Yours,

—-------------------------------------
Md Mehedi Hasan
ECE Department
North South University

—-------------------------------------
Ahmed Ajmine Nehal
ECE Department
North South University

—-------------------------------------
Farhana Elias
ECE Department
North South University

APPROVAL

Md Mehedi Hasan (ID # 2022107642), Ahmed Ajmine Nehal (ID # 2013090042), and Farhana Elias (ID # 2013820042) from the Electrical and Computer Engineering Department of North South University, have worked on the Senior Design Project titled “Named Entity and Relation Extraction Aware Biomedical Extractive Question Answering using Transformers with BiLSTM” under the supervision of Dr. Mohammad Rashedur Rahman, which has been accepted as satisfactory and satisfied the requirement for the degree of Bachelors of Science in Engineering.

Supervisor’s Signature

…………………………………….

Dr. Mohammad Rashedur Rahman

Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

Chairman’s Signature

…………………………………….

Dr. Rajesh Palit

Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

DECLARATION

This is to certify that this work is entirely original to us. No portion of this work has ever been submitted elsewhere, in whole or in part, to be considered for a different degree or diploma. All information pertaining to the project will be kept private and shall not be divulged without the prior authorization of the project supervisor. All pertinent prior works cited and acknowledged in this report are relevant. The supervisor's declared anti-plagiarism guideline has been followed.

1. Md Mehedi Hasan

—-------------------------------------

2. Ahmed Ajmine Nehal

—-------------------------------------

3. Farhana Elias

—-------------------------------------

ACKNOWLEDGEMENTS

Dr. Mohammad Rashedur Rahman, Professor in the Department of Electrical and Computer Engineering at North South University in Bangladesh, is greatly appreciated by the authors for his precise instructions, support, and guidance about the experiments, research, and theoretical studies conducted during the current project and in the writing of the current report.

Furthermore, The Department of Electrical and Computer Engineering at North South University in Bangladesh is also appreciated by the authors for making this research possible. The authors would also like to express their gratitude to their family members for their unending sacrifices and unwavering support.

ABSTRACT

Transformers and BiLSTM-Based Ensemble Modeling for Entity and Relation Aware Biomedical Extractive Question Answering

In the realm of natural language processing, extractive question answering (EQA) from a given context paragraph is a fascinating task that has attracted the attention of researchers due to its practical applications and promising future across diverse fields. While EQA has flourished in general domains with abundant resources, the biomedical realm presents an exciting frontier for innovation. This emerging landscape offers a unique opportunity to contribute significantly and accelerate progress. This study presents a novel approach to biomedical factoid-based extractive Question Answering, addressing the complexities of medical language. We leverage transformer-based pre-trained models, enhancing them with Named Entity Recognition, Relation Extraction, and BiLSTM layers. Our method, culminating in ensemble learning, achieves a 5.45% performance boost. The resulting system not only improves answer accuracy but also uncovers intricate biomedical relationships, providing deeper insights for researchers and clinicians. By bridging the gap between complex medical texts and precise information retrieval, our work contributes significantly to advancing biomedical knowledge extraction. Our study progressed through a systematic evaluation of six transformer encoder models, ultimately selecting Bio+ClinicalBERT as the foundation for our approach. We developed three increasingly sophisticated BEQA models, culminating in Bio+ClinicalBERT_NER_RE_BiLSTM. This final model excelled in performance, achieving F1 91.69, Exact Match 88.35, and Lenient Accuracy 0.84. These results demonstrate our model's capability to navigate complex biomedical language, extract precise information, and significantly advance the field of biomedical question answering.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Background and Motivation

A growing number of researchers in the fields of Natural Language Processing (NLP) [1]and Information Retrieval (IR) [2], [3] are interested in studying Question-Answering (QA) systems. In NLP, QA research interest has increased primarily due to improvements in Deep Learning efficacy [4]. The process of locating answers to questions within the given context is known as Extractive Question Answering (EQA) [5], [6]. Using annotated training data, EQA extracts text from documents to answer user questions. EQA types include yes/no, factoid, list, and summarize, with factoid-type QA being particularly important due to its practical applications in information retrieval and decision support systems [7]. Two popular benchmarking datasets for EQA are SQuAD (Stanford Question Answering Dataset) [8] and BioASQ. SQuAD focuses on general domain questions, while BioASQ [9] specifically targets biomedical literature. These datasets are connected in that they both aim to evaluate and improve QA systems, with BioASQ extending the concepts of SQuAD to the specialized field of biomedicine. Biomedical Extractive Question Answering (BEQA) is a subset of EQA that focuses on extracting answers from biomedical texts, addressing the unique challenges of medical terminology and complex scientific concepts. These challenges inspired and motivated us to develop a QA model to effectively extract precise answers and make technological advancements in the biomedical field.

BEQA faces challenges like domain-specific terminology, complex scientific concepts, and the need for up-to-date knowledge [10]. Improvements can be made in handling biomedical jargon, interpreting context-dependent information, and integrating external knowledge bases. The scope for advancement includes developing models that understand biomedical relationships, improving entity recognition in specialized texts, and creating diverse biomedical QA datasets for better training and evaluation. Due to the sensitivity of domain-specific challenges, sophisticated solutions are required. Deep learning techniques such as fine-tuning, transfer learning, adding BiLSTM layers, and ensemble learning with bootstrap sampling and majority voting can be employed. These approaches enhance the system's ability to handle complex biomedical terminology and concepts. Utilizing the latest datasets for training and evaluation is crucial for maintaining up-to-date knowledge and improving performance. These solutions collectively aim to increase the accuracy and reliability of BEQA systems in interpreting and extracting relevant information from specialized biomedical texts.

Transformers have significantly improved BEQA by handling complex language patterns and long-range dependencies [11]. Pre-trained Transformers-Encoder models like BERT, and BioBERT excel in BEQA tasks, capturing contextual information, handling domain-specific vocabulary, and enhancing performance with fine-tuning. Despite the advancements in BEQA through transformer models, there is a gap in fully integrating combined deep learning approaches for factoid-type questions [12]. Traditional methods and individual deep learning techniques have been explored, but comprehensive integration is needed for better understanding complex biomedical questions, connecting them to relevant contexts, and extracting precise factoid answers. Unlike other architectural models, transformer models excel in this domain due to their superior ability to capture long-range dependencies and contextual nuances in biomedical texts. However, There is potential for further improvement by combining transformers with other deep learning strategies to enhance question comprehension, context-answer relationship modeling, and accurate factoid extraction in BEQA systems. To this end, our BEQA system aims to improve question comprehension, context-answer relationship modeling, and factoid extraction in biomedical texts by integrating deep learning techniques and models based on the transformer architecture.

1.2 Purpose and Goal of the Project

In our work, we used six pre-trained models—BERT, BioBERT, PubMedBERT, ClinicalBERT, Bio+ClinicalBERT, and BioLinkBERT—to finetune on the BioASQ 4b factoid training set for question-answer downstream tasks. Then, comparing them by the highest outcome of evaluation metrics—F1 and EM—we chose the Bio+ClinicalBERT pre-trained model for the next steps. As the application of transfer learning has garnered a lot of interest in the biomedical field and there is usually little data available for training in biomedical quality assurance tasks, transfer learning is a good option for incorporating additional external knowledge. Therefore, we used Named Entity Recognition (NER) and Relation Extraction (RE) tasks as Transfer Learning techniques in the pre-trained Bio+ClinicalBERT model. Next, we added the BiLSTM layer and then fine-tuned it on the BioASQ 9b factoid training set. We then used the ensemble learning technique to find a tuned model further. Finally, we developed our three BEQA models—Bio+ClinicalBERT_NER QA Finetuned, Bio+ClinicalBERT_NER_RE QA Finetuned, and Bio+ClinicalBERT_NER_RE_BiLSTM QA Finetuned—where we found our best-performing final model, Bio+ClinicalBERT_NER_RE_BiLSTM QA Finetuned, which achieves state-of-the-art results.

This paper uses advanced pre-trained transformer models and deep learning techniques to present an advanced Biomedical Extractive Question Answering (BEQA) system. The system excels in handling complex biomedical terminology and extracting precise factoid answers. The best-performing model, Bio+ClinicalBERT_NER_RE_BiLSTM QA Finetuned, is a significant advancement in BEQA technology. This study contributes to understanding BEQA capabilities and lays the groundwork for developing influential tools for researchers and clinicians, potentially revolutionizing information retrieval and decision support in healthcare and biomedicine. In conclusion, the following summarizes our main contributions to this study:

- Development of an advanced BEQA model capable of generating accurate answers to biomedical questions from given passages.
- Fine-tuning of six biomedical-specific pre-trained models (BERT, BioBERT, PubMedBERT, ClinicalBERT, Bio+ClinicalBERT, and BioLinkBERT) for in-depth evaluation on the BioASQ dataset in question-answering tasks.
- Integration of NER and RE to enhance information extraction and entity relationship identification within biomedical texts.
- Incorporation of BiLSTM layers and ensemble learning techniques to further boost model performance.
- Creation of our top-performing model, Bio+ClinicalBERT_NER_RE_BiLSTM QA Finetuned, which achieves state-of-the-art results in biomedical extractive question answering.

1.3 Organization of the Report

The paper is structured as follows: Section 2 reviews relevant literature, emphasizing the shortcomings in biomedical question answering for BERT-based and other models and how our study overcomes them. Section 3 describes the technique that uses Bio+ClinicalBERT, BioASQ data, ensemble learning, bootstrap sampling, and the transfer of NER and RE knowledge. It also presents how three BEQA models are developed, the last of which is Bio+ClinicalBert_NER_RE_BiLSTM. Section 4 details the experimental setup, metrics, model comparisons, and evaluations. Section 5 discusses the findings, and Section 6 concludes with future directions.

Chapter 2 Research Literature Review

Biomedical Extractive Question Answering (BEQA) has gained attention for accurate and efficient extraction of information from complex biomedical texts. Research has explored integrating Named Entity Recognition and Relation Extraction tasks, transformer-based models, and hybrid learning approaches combining deep learning techniques like BiLSTM and ensemble learning. These approaches have improved the handling of domain-specific terminology and enhanced BEQA system performance.

2.1 Existing Research

2.1.1 Biomedical Question Answering

A framework for answering questions with perfect and exact answers has been provided in a number of recent research, and state-of-the-art results on questions with perfect answers on the BioASQ dataset have been obtained [13]. They use LeToR ranking models to respond to factual or list-based queries, neural entailment models for precise replies, and a unique embedding transformation method for yes/no questions. They enhance the IR part of extractive summarization for optimal responses. The generated summary answer's human readability still needs a great deal of improvement, even if this greatly raises ROUGE scores.

For yes/no, factoid, and list questions [14], SemBioNLQA can provide specific answers (e.g., "yes," "no," a biomedical entity, etc.) as well as paragraph-sized ideal replies (summaries of pertinent information). On the other hand, it only returns optimal responses to summary queries.

Meanwhile, the framework LIQUID—which automatically creates a list of QA datasets from unlabeled corpora—was presented to address the issue of data scarcity [15]. On five benchmark datasets, the synthetic data significantly enhanced the performance of the existing supervised models. Here, the impact of every LIQUID component was carefully examined, and quantitative and qualitative data were produced.

2.1.2 BioMedical QA with Name Entity Recognition

Biomedical Question Answering (QA) systems aim to extract relevant information from vast amounts of biomedical literature in response to user queries. Named Entity Recognition plays a crucial role in this process by identifying and classifying key entities within the text, such as genes, proteins, diseases, and drugs. Accurate NER serves as a foundation for downstream tasks like relation extraction and ultimately improves the overall performance of BioMedical QA systems.

Several recent studies have explored the integration of NER techniques to enhance BioMedical QA. The impact of NER on a question answering system focused on pharmacological substances and compounds [16]. Their findings demonstrated that incorporating a state-of-the-art NER model significantly improved the system's ability to answer questions involving these entities.

Deep learning architectures have developed robust NER models for biomedical literature. A Bidirectional Long Short-Term Memory (BiLSTM) network for better detection of biomedical entities [17] and Multi-task learning and pre-trained language models work for excellent accuracy [18]. Their model effectively captured contextual information from biomedical text and achieved high accuracy in recognizing diverse biomedical entities.

These studies highlight the significant contribution of NER to BioMedical QA. By pinpointing relevant entities within the text, NER systems enable QA systems to focus on crucial information and provide more accurate and informative answers to user queries.

2.1.3 Relation Extraction in BioMedical QA

BioMedical Question Answering (QA) systems not only need to identify relevant entities within the text, but also comprehend the connections between these things to provide comprehensive and informative answers. Relation extraction (RE) plays a vital role in this process by uncovering the semantic links between entities, such as protein-protein interactions, drug-disease associations, and gene regulatory pathways. By capturing these relationships, RE empowers BioMedical QA systems to reason over the extracted information and deliver more insightful responses to complex user queries.

Recent research has highlighted the importance of RE for enhancing BioMedical QA performance. A deep learning-based approach for RE that utilizes convolutional neural networks (CNNs) to capture local patterns and recurrent neural networks (RNNs) to model long-range dependencies within sentences [19]. Their system achieved promising results in extracting various biomedical relations from text.

Further advancements have been made in exploiting external knowledge resources to improve RE accuracy. A method that integrates knowledge graphs with a neural RE model [20]. This approach leverages pre-existing knowledge about biomedical entities and relationships in order to improve the model's performance to extract novel relations from unseen text.

These studies demonstrate the significant influence of RE on BioMedical QA performance. By extracting relationships between identified entities, RE empowers QA systems to move beyond simple entity recognition and provide a deeper understanding of the underlying biological processes. This capability is crucial for answering complex biomedical queries that require reasoning over interconnected entities and relationships.

2.1.4 Transformer and BiLSTM Architectures for BioMedical QA

This work explores the effectiveness of Transformer and Bi-directional Long Short-Term Memory (BiLSTM) architectures in Question Answering (QA) tasks. It compares and contrasts the ability of these deep learning models to convey contextual relationships and long-range dependencies in text for improved answer extraction.

In the context of the MEDIQA challenge, the LasigeBioTM team proposed an approach that explored a common Transformer-based architecture for each task, utilizing the given training data to adjust the same pre-trained weights for every activity [21]. The group used NER, a named entity recognition tool, to enhance the question and answer texts and supplemented the training data with additional datasets.

In the context of the BioASQ10B challenge, The use of BioM-Transformers models, an adaptation of both ELECTRA and ALBERT models to the biomedical domain [22]. The authors extended their investigation of biomedical QA models with BioM-Transformers by combining the BioASQ10B-Factoid training set with the List training set to overcome its small size, and by performing a grid search for hyperparameters.

Several studies have explored the application of BiLSTM-CRF models for NER in the context of online health communities. For instance, A BiLSTM-CRF model for NER in a diabetes-focused online community achieved promising results in identifying medical entities [23]. Their work highlights the potential of this approach for extracting valuable medical information from doctor-patient interactions. Bidirectional Long Short-Term Memory (BiLSTM) networks are a popular choice because they can identify long-range dependencies within text data. Conditional Random Fields (CRFs) are often employed in conjunction with BiLSTMs to model the sequential relationships between entities and improve recognition accuracy.

2.1.5 Biomedical QA with Hybrid Learning Approach

Utilizing Transfer Learning and Ensemble Learning techniques to improve the performance of Question Answering (QA) systems [24]. The work examines how combining multiple diverse QA models (ensemble learning) can lead to more robust and accurate answer retrieval.

In the realm of biomedical question answering with Transformer ensembles, several notable works have contributed to advancing the field. An ensemble learning pipeline that leverages Large Language Models (LLMs) to improve medical question-answering tasks' precision and dependability [24]. Their study focused on improving performance across diverse medical QA datasets, including PubMedQA, MedQA-USMLE, and MedMCQA, by employing two primary ensemble methods: Boosting-based weighted majority vote ensemble and Cluster-based Dynamic Model Selection. The results demonstrated superior performance compared to individual LLMs, showcasing accuracies of 35.84\%, 96.21\%, and 37.26\% for MedMCQA, PubMedQA, and MedQA-USMLE, respectively, with the Majority Weighted Vote method.

A transfer learning-based sentiment-aware model named SentiMedQAer for biomedical question answering (QA) [25] aims to address limitations in existing biomedical QA datasets, particularly their small size and the prevalence of factoid questions. The study aims to enhance the accuracy and reliability of medical question-answering (QA) tasks by developing an ensemble learning pipeline that utilizes state-of-the-art large language models [24]. The goal is to improve performance on diverse medical QA datasets.

Various techniques for data augmentation, ensemble learning, model training, and pre-processing for biomedical QA, using advanced pre-trained models like GPT-4 and BioLinkBERT [26]. Their study achieved top rankings in each of the four batches of BioASQ Task 11b-Phase B questions that are yes/no in nature, one of the four batches that are factoid in nature, and two of the four batches that are list-type questions.

Another approach to transfer learning for biomedical factoid question answering is named entity aware transfer learning [27]. This approach involves using named entities, such as genes, drugs, and diseases, to guide the transfer learning process. For example, they proposed a named entity aware transfer learning method that uses a pre-trained language model to extract features from biomedical literature, with a focus on named entities. The authors reported that their approach achieved significant improvements over traditional transfer learning methods.

2.2 Existing Research Limitations

Table 1. Related Works in Biomedical QA, NER, and RE

Ref No	QA Form	Research	Model Used	Dataset	Findings	Limitations/Research Gap
[13]	Biomedical QA	BioAMA: End to End Biomedical QA System	IR-based, NLI-based using Hierarchical CNN, Two-stage using NER taggers and ranking algorithms	BioASQ 5b	ROUGE-2: 0.72, ROUGE-SU4: 0.71, NLI framework accuracy: 65.6%	Inadequate training data for NLI model, needs embedding projection technique, continued refinement and optimization for different question types
[15]	List QA	LIQUID: Framework for List QA Dataset Generation	Summarization Model: BARTbase, QG Model: Trained on SQuAD, QA Model: Trained on	MultiSpanQA, Quoref, BioASQ 7b, 8b, 9b	Improved F1 scores: MultiSpanQA: 5.0, Quoref: 1.9, BioASQ datasets: 2.8	Efficiency issues, quality of synthetic data, need optimization for iterative filtering and answer expansion stages

			SQuAD, NER Models			
[16]	Named Entity Recognition	PharmaCoNER: Pharmacological Substances, Compounds, and Proteins Named Entity Recognition Track	Summarization Model: BART, NER Model, QG Model: Trained on SQuAD, QA Model: Trained on SQuAD	MultiSpanQA, Quoref, BioASQ 7b, 8b, 9b	Improved performance with synthetic data, enhanced quality of generated QA pairs	Reliance on high-quality summarization and NER models, iterative filtering process needs optimization
[17]	Named Entity Recognition	Fusion Attention-Based BiLSTM-CRF for Biomedical Named Entity Recognition	BiLSTM, Attention Mechanism, CRF	JNLPBA, BC2GM	F1-score of 73.50 on JNLPBA, attention mechanism prevents significant information loss	Need for validation on other biomedical corpora, improvement in automatic feature learning, optimization of attention mechanism

[18]	Named Entity Recognition	BioNER with Combined Feature Attention and Fully-shared Multi-task Learning	Vanilla BioBERT, BioKMNER, BioBERT-TWA, BioBERT-CFA, Fully-shared MTL	BC2GM, JNLPBA, BC5CDR-Disease, NCBI-Disease, Linnaeus, Species-800, BC5CDR-Chemical	BioBERT-CFA model achieved competitive performance, MTL model improved BioNER tasks by leveraging cross-type information	Further incorporation of syntactic features, need for generalization across different biomedical entities
[20]	Relation Extraction	Biomedical Relation Extraction with Knowledge Graph-Based Recommendations	Ontology Embedding Layer, LSTM, TUP Model	BC2GM, JNLPBA, BC5CDR-Disease, NCBI-Disease, Linnaeus, Species-800, BC5CDR-Chemical	Improved performance with single-task and multi-task models, better performance with ensemble methods	Impact of different syntactic information on multi-task model performance, integration of more diverse biomedical data
[21]	Biomedical QA	LasigeBioTM at MEDIQA 2019: Biomedical QA using Bidirectional Transformers and NER	BERT, BioBERT	NLI training set, RQE training set, QA training set, MedQuAD Cancer-Gov dataset	High accuracy on NLI task, improved RQE and QA performance with NER augmentation	Drop in accuracy from development to test set, ranking issues in detecting correct answers
[22]	Biomedical QA	Exploring Biomedical QA with BioM-Transformers at BioASQ10B Challenge	BioM-Transformers: Adaptations of ELECTRA and ALBERT	BioASQ10B training and test datasets, BioASQ classification dataset	Improved performance by combining training sets, extensive hyperparameter tuning, varied	Technical and time constraints, reproducibility concerns with Transformer-based models

					performance across batches	
[24]	Medical QA	One LLM is not Enough: Harnessing the Power of Ensemble Learning for Medical QA	Boosting-based Weighted Majority Vote Ensemble, Cluster-based Dynamic Model Selection	MedMCQA, PubMedQA, MedQA-USMLE	Superior performance with ensemble methods, significant accuracy improvement on medical QA datasets	Challenges in ensembling LLMs, cost associated with training large models, need for further exploration in structured QA tasks
[26]	Biomedical QA	Exploring Approaches to Answer Biomedical Questions: From Pre-processing to GPT-4	BioLinkBERT, GPT-4	BioASQ-10b, BioASQ-11b, SQuAD, LIQUID	High rank in BioASQ Task 11b, BioLinkBERT outperformed existing models, GPT-4 effective in list-type questions	Limited exploration of data augmentation for yes/no questions, further exploration of GPT models for different question types
[27]	Factoid QA	Named Entity Aware Transfer Learning for Biomedical Factoid QA	BioBERT, BiLSTM, Ensemble Method (Bagging)	BioASQ 6b, 7b	Improved performance with fine-tuned BioBERT and BiLSTM, effective ensemble approach	Complexity of biomedical named entities, data imbalance, challenges in learning sentence embeddings

[28]	Biomedical QA	Improving Biomedical QA with Sentence-based Ranking at BioASQ-11b	Sentence embedding cosine similarity ranking, BioM-ELECTRA model	BioASQ 2023 11b dataset, BioASQ 7b, PubMedQA	Improved accuracy for yes/no and factoid questions, sentence ranking step enhanced QA system performance	Limited improvement in ideal answer generation performance, dependency on extractive QA approach
[29]	Relation Extraction	Biomedical Relation Extraction with Entity Type Markers and Relation-specific QA	Span QA-based Relation Extraction, Binary QA-based Relation Extraction	DrugProt dataset	Better performance with Binary QA-based method, entity type markers improved model effectiveness	Focused on DrugProt dataset, scalability issues with manual creation of question templates
[25]	Sentiment-aware QA	Transfer Learning-based Sentiment-aware Model for Biomedical QA	BioBERT, T5, RoBERTa, XGBoost	PubMedQA dataset	Outperformed SOTA by 15.83%, integration of sentiment information significantly enhanced performance	Generalizability limited to PubMedQA dataset, further exploration needed for other question types

2.3 Overcomes of Existing Research Limitations

Our research uniquely combines multiple advanced techniques to address key limitations in existing biomedical question answering systems. Unlike previous approaches that rely solely on single models or limited combinations (e.g., [21], [22]), we integrate NER, RE, and BiLSTM layers with the Bio+ClinicalBERT model. This comprehensive approach mitigates the need for embedding projection and optimization for different question types, as seen in [13]. By incorporating NER and RE, we enhance the model's ability to handle domain-specific terminology and complex relationships, addressing limitations noted in [16] and [18]. Our use of ensemble learning with bootstrap sampling and majority voting improves upon the single-model approaches in [20] and [24], leading to more robust performance across diverse biomedical questions. Furthermore, our method achieves high accuracy without relying on extensive hyperparameter tuning or large-scale datasets, which were limitations in [22] and [26]. By addressing these gaps, our research advances the field of biomedical extractive question answering, offering improved accuracy and reliability in processing complex biomedical information.

Chapter 3 Methodology

3.1 Data and resources

3.1.1 Dataset Breakdown

The BioASQ Question Answering (BioASQ-QA) dataset is an invaluable resource for academics working on factoid question answering systems in the biomedical area. Unlike many question-answering datasets, BioASQ-QA focuses on simulating the information requirements of actual biomedical researchers. Each year, the dataset is expanded by approximately 500 new questions posed by specialists, ensuring its relevance to current research trends and concerns [9]. The BioASQ training Phase B dataset divides questions into four kinds, each with a unique strategy for a question answering system. In our effort, we concentrated on factoids. Factoid inquiries seek a brief, objectively verifiable answer within the context of the biomedical literature supplied. The goal is to develop a system that can accurately retrieve and present these factual answers. By focusing on factoids, we aim to improve the efficiency and accuracy of question answering in the biomedical domain.

A. Question Answering Dataset

Initially, we used the factoid type of bioasq-4b dataset of "BioASQ.org" for training and evaluating the performance of several Extractive Question Answering Models to determine the best model. The following dataset's train and test splits have 327 unique question rows with 3266 question-answer pairs. In our final effort, we employed the recent version of the BioASQ question answering dataset, bioasq-9b, which has a higher amount of data. The bioasq-9b dataset contains a total of 1092 unique question rows with 5447 question-answer pairs, providing a larger and more diverse set of data for training and evaluating the models. We keep the train and test set splitting into 80\% and 20\% of dataset 4b, and 9b. This allowed us to further improve the performance and accuracy of our BEQA Models.

Table 2. Statistics of BioASQ Phase B Dataset

Version	Train Samples	Post-processed Question-passage pairs
4b	327	3266
9b	1092	5447

The dataset contains 4 features named:

1. id
2. question
3. context
4. answers
    1. text
    2. answer_start

Table 3. An instance of BioASQ Phase B dataset

id(string)	question(string)	context(string)	answers(sequence)
"53148a07dae131f847000002_001"	"**Which drug is considered as the first line treatment of fibromyalgia?**"	"Pregabalin: a review of its use in fibromyalgia. Oral pregabalin, a calcium channel alpha(2)delta-subunit ligand with analgesic, anxiolytic and antiepileptic activity, has shown efficacy in the treatment of fibromyalgia. It has a multidimensional effect in the treatment of this complex condition, and is associated with rapid and clinically significant improvements in several outcome measures relating to core symptoms of the syndrome, including pain and sleep, in patients with long-standing fibromyalgia. Pregabalin treatment is also associated with improvements in the overall health status of these patients. The beneficial effects of pregabalin are durable in patients with an initial response to the drug. The most common adverse events associated with the drug are dizziness and somnolence, which are generally mild to moderate in intensity and are tolerated by many patients. **Pregabalin is, therefore, a valuable option in the first-line treatment of patients with fibromyalgia**."	"text": "**Pregabalin**", "answer_start": 886

In the provided example in Table 3,the question is "Which drug is considered as the first line treatment of fibromyalgia?" The context offers detailed information about the drug Pregabalin, which is identified as the answer. The dataset includes both the answer text and its position in the context (e.g., "answer_start": 886). This structure allows QA models to directly locate and retrieve relevant biomedical information, making it highly effective for applications in healthcare and clinical research. By utilizing this method, researchers and healthcare professionals can quickly access critical data to inform decision-making and treatment strategies.

B. Named Entity Recognition Dataset

NER in the biomedical field extracts valuable entities from unstructured texts like research articles or clinical notes. It works by tagging words (tokens) with labels (tags) indicating their entity types or whether they are outside (O) of any entity class. This process is crucial for organizing large-scale biomedical data, enabling improved information retrieval, and supporting medical research. The relation between tokens and tags shows how specific terms, like diseases or genes, are recognized and classified by these systems. There are different types of datasets that aim to identify and classify entities such as diseases, chemicals, genes, and species.

Table 4. Statistics of Ner Datasets

Dataset	Source	Entity Type	Quantity
BC2GM	PubMed	Gene and proteins	20,000
BC4CHEMD	PubMed	Chemical entities	100,000
BC5CDR-disease	PubMed	Diseases, chemicals, and their relationships	6,000
JNLPBA	GENIA corpus	Biomedical entities (e.g., genes, proteins)	2,000
Linnaeus	Scientific publications	Species names and other taxonomic entities	100
NCBI-disease	National Center for Biotechnology Information (NCBI)	Diseases (linked to standardized medical terms)	6,000
S800	PubMed	Species and taxonomic entities	800

The NER datasets vary by entity type and data sources. BC2GM, BC4CHEMD, BC5CDR-disease, and JNLPBA are commonly used for identifying biomedical entities like genes, chemicals, and diseases. Linnaeus and S800 focus on species names, with Linnaeus providing higher-quality annotations. NCBI-disease is designed specifically to extract disease mentions and link them to standardized vocabularies, making it valuable for tasks requiring consistent medical terminology.

Table 5. An instance of Ner Dataset

Token	**Tag**
Identification	O
of	O
APC2	O
,	O
a	O
homologue	O
of	O
the	O
adenomatous	B
polyposis	I
coli	I
tumor	I
suppressor	O
.	O

In the given NER dataset instance in Table 5 ,tokens are individual words or punctuation marks, while tags represent the entity categories. The "B" tag indicates the beginning of an entity, "I" indicates the continuation of that entity, and "O" signifies that a token does not belong to any entity. For example, in the instance provided, "adenomatous" is tagged as "B" (beginning of an entity), followed by "polyposis" tagged as "I" (continuation), representing a multi-token biomedical entity. The tokens "adenomatous," "polyposis," "coli," and "tumor" collectively indicate a biomedical entity, specifically a disease. Therefore, these tokens belong to the category of diseases in a Named Entity Recognition (NER) dataset.

C. Relation Extraction Dataset

Relation extraction (RE) in biomedicine identifies semantic relationships between entities like genes and diseases. It works by analyzing sentences and assigning labels to indicate whether specific relationships exist. In an RE dataset, sentences are annotated to indicate if a specific semantic relationship exists between entities like genes and diseases. Entities are represented using placeholders like @GENES and @DISEASES. The labels (1 or 0) show whether a relevant relation is detected in the sentence. The RE model learns to identify such relationships based on patterns, phrases, and entity interactions within the text.

Table 6. Statistics of RE Datasets

Dataset	Source	Entity Type	Quantity
EuADR	Biomedical abstracts	Drug-condition relationships	1,000 abstracts
GAD	Genetic Association Database	Gene-disease associations	200,000 records

The EuADR dataset focuses on drug-condition relationships and is annotated from biomedical literature, while the GAD dataset specializes in gene-disease associations, with a large collection of annotated genetic association studies. These two datasets are valuable for advancing biomedical research and supporting the development of applications like drug safety monitoring and personalized medicine.

Table 7. Instances of RE dataset

Sentence	Label
Common polymorphisms in the genes @GENES and LOC387715 are **independently related to** @DISEASES progression after adjustment for other known AMD risk factors.	1
A large @DISEASES of exons 9 and 10 of @GENES confers an **increased risk** of prostate cancer in Polish men.	0

The first sentence in the Instances of RE dataset in Table 7 indicates a relationship between a gene and a disease. The key phrase "independently related to" suggests a direct association between the gene (@GENES) and disease (@DISEASES) mentioned, hence the label "1," indicating a positive relationship. The last sentence mentions gene regions (exons) but does not establish a direct relationship between the entities labeled as @GENES and @DISEASES. The phrase "increased risk" could imply a potential relationship, but the structure and context imply that it is indirect or hypothetical, leading to the label "0," indicating no direct relationship.

3.1.2 Datasets Collaboration

The collaboration between QA, NER, and RE tasks in biomedical NLP provides a comprehensive approach for extracting and understanding complex information from text. This approach enhances biomedical question answering by combining NER, RE, and QA. NER identifies key entities, RE maps their relationships, and QA extracts precise answers. Together, they improve accuracy and depth, aiding research and clinical decisions.

Table 8. Collaboration of NER,RE and QA tasks to understand how the knowledge of NER and RE works for extracting answers of questions from the context accurately

Question: What gene is associated with cystic fibrosis?	
Context: Cystic fibrosis is a genetic disorder that affects the lungs, pancreas, and other organs. It is caused by mutations in the CFTR gene, which codes for a protein involved in salt and water movement across cell membranes.	
NER:	**Entities**: • Disease: Cystic fibrosis • Gene: CFTR
RE:	**Relations**: • [Cystic fibrosis @DISEASE$] is_caused_by [CFTR @GENE$]
Answer: CFTR.	

In the collaboration Table 8, NER identifies key entities like diseases and genes in the context, for instance, tagging “Cystic fibrosis” as a disease and “CFTR” as a gene. RE determines the relationship between these entities, such as identifying that "Cystic fibrosis" is caused by "CFTR". QA combines both NER and RE outputs to answer specific questions directly, like “What gene is associated with cystic fibrosis?” By integrating the entities and relations detected, QA provides the precise answer: “CFTR”.

3.1.3 Data Preparation and Pre-processing

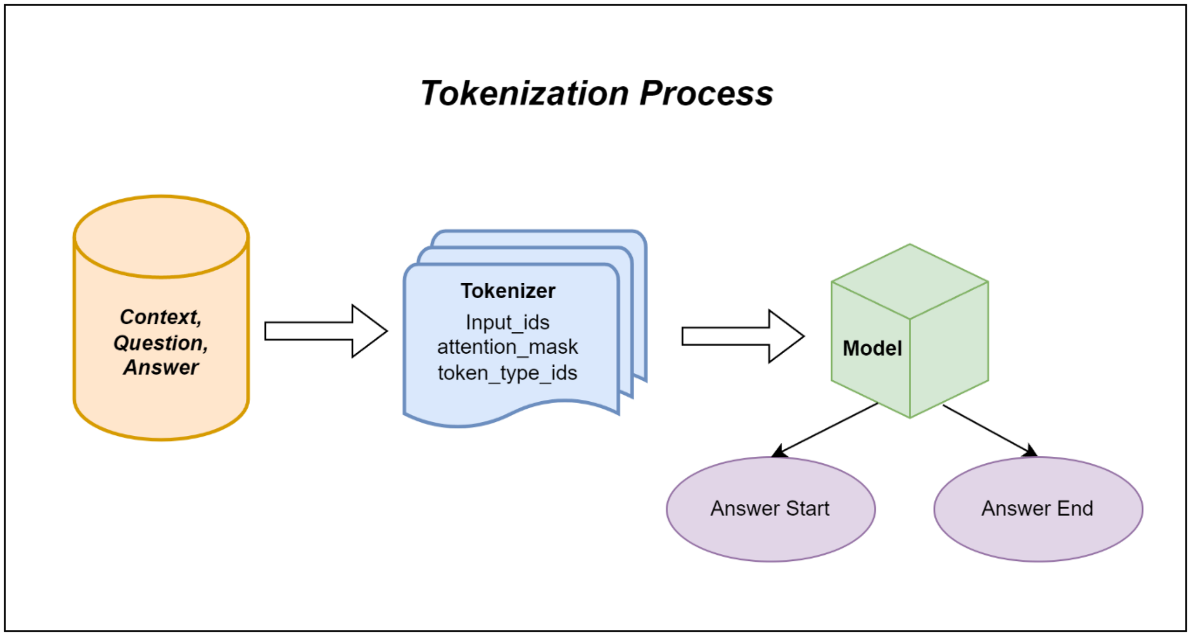


Figure 1. Tokenizer Processing Architecture

Standard text preprocessing techniques, including tokenization, stemming, and stop-word removal, are applied to ensure compatibility with downstream QA models. Each model's specific tokenizer is used to process the dataset, preserving key linguistic nuances that generic tokenizers might miss. This refined dataset is then utilized for training and evaluating the QA models, promoting a standardized and unbiased approach to question answering. This approach ultimately leads to more accurate and reliable QA model performance.

Table 9. A Tokenizer processing Example of Question and Answer

Tokenizer: emilyalsentzer/Bio_ClinicalBERT	
Question	**Tokenized (WordPeice Tokenization)**
Which hormone abnormalities are characteristic to Pendred syndrome?	'[CLS]', 'Which', 'hormone', 'abnormal', '##ities', 'are', 'characteristic', 'to', 'Pen', '##dre', '##d', 'syndrome', '?', '[SEP]'
Context	**Tokenized (WordPeice Tokenization)**
DOCA sensitive pendrin expression in kidney, heart, lung and thyroid tissues.	'[CLS]', 'D', '##OC', '##A', 'sensitive', 'pen', '##dr', '##in', 'expression', 'in', 'kidney', ',', 'heart', ',', 'lung', 'and', 'thy', '##roid', 'tissues', '.', '[SEP]'

The tokenizer used the example in Table 9, is Bio_ClinicalBERT, specifically designed for biomedical text processing. It employs WordPiece tokenization, which breaks down words into subword units like "abnormalities" into "abnormal" and "\#\#ities". The tokenizer's context-aware segmentation leads to more accurate question answering and relationship extraction in biomedical datasets, improving overall model performance in clinical NLP tasks. By preserving subword information and enabling finer-grained segmentation of biomedical terms, the tokenizer improves the model’s understanding of domain-specific language, leading to enhanced performance on clinical datasets. This practical tokenization approach ensures that the nuances of medical terminology are captured, which is crucial for delivering accurate and contextually relevant results in biomedical NLP applications.

3.2 Model Selection

In this study, we evaluated the performance of six transformer models—BERT, BioBERT, ClinicalBERT, BioClinicalBERT, PubMedBERT, and BioLinkBERT—focusing on fine-tuning for an extractive Question-Answering (QA) task. A consistent architecture across all models ensures a fair performance comparison, with BERT serving as a benchmark. The transformer encoder, central to these models, operates using a combination of a fully connected feed-forward network and multi-head self-attention [30]. Key components of the encoder include Input Embedding, where tokens are transformed into dense vector representations capturing semantic information, and Positional Encoding, which is added to embeddings to help the model understand the order of words, since transformers lack inherent sequential awareness. Self-Attention allows each token to compute attention scores with every other token through queries, keys, and values, capturing relationships between words regardless of their position in the text. After self-attention, the token representations are passed through a fully connected Feed-Forward Network, enhancing the model's ability to learn non-linear patterns. This architecture allows for effective learning from complex biomedical texts and ensures robust performance in QA tasks.

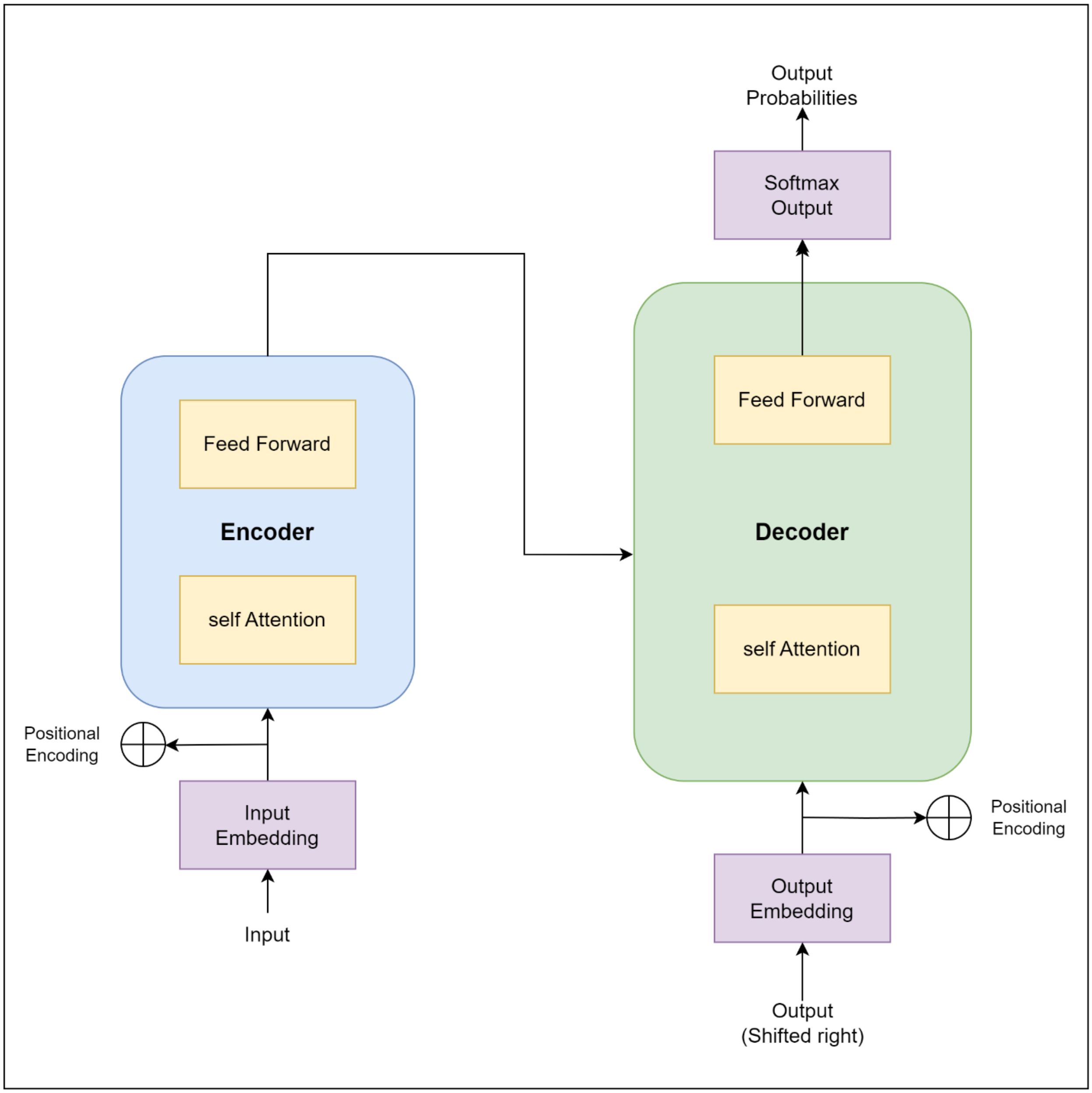


Figure 2. Transformers Architecture

Brief descriptions of the models are provided as follows:

3.2.1 Transformers Pre-trained Models

A. BERT

BERT as joint conditioning on both the left and right context in all layers to pre-train deep bidirectional representations from the unlabeled text [31]. Therefore, without requiring significant task-specific architecture changes, the pre-trained BERT model may be improved with just one extra output layer to produce state-of-the-art models for a variety of tasks, including question answering and language inference. This makes it a versatile and powerful tool for researchers and practitioners in the field of NLP.

B. BioBERT

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is a unique language model developed on large medical text collections [32]. It learned a lot from PubMed and PMC journals. BioBERT has almost the same architecture for different tasks, and it works better than BERT and other top models in various tasks related to understanding and extracting information from biomedical texts.

C. ClinicalBERT

A pre-trained language representation model called ClinicalBERT was created especially for clinical literature in the medical field. It is the BERT's expansion. To capture domain-specific semantics and contextual information found in medical literature, electronic health records (EHRs), and other clinical documents, ClinicalBERT is trained on large-scale clinical text data. ClinicalBERT is better able to comprehend the subtleties and linguistic quirks unique to the healthcare industry thanks to this fine-tuning of medical data [33].

D. Bio+ClinicalBERT

Bio+Clinical BERT is a powerful language model specifically designed for the healthcare domain. It builds upon the foundation of two existing models: BioBERT and regular BERT. Bio+Clinical BERT combines the strengths of both models. It starts with the pre-trained weights of BioBERT and then further refines them by training on a vast collection of electronic health Records from MIMIC III, a database of ICU patient notes [34].

E. PubMedBERT

PubMedBERT model developed by Microsoft. This model underwent meticulous pretraining from the ground up, utilizing a diverse dataset comprising complete-text publications from

PubMedCentral and abstracts from PubMed. This comprehensive training approach aimed to establish a robust understanding of biomedical language nuances. PubMedBERT has demonstrated modern performance on a range of biomedical natural language processing (NLP) tasks, such as extractive question answering, underscoring its importance in the area [35].

F. BioLinkBERT

BioLinkBERT is a language model designed for biomedical knowledge graph completion tasks, combining BERT's contextual information capture and knowledge graph embeddings to enhance performance in tasks like link prediction and entity typing by learning over biomedical entities [36].

Table 10. Corpora of Pre-trained Models

Pre-TrainedModels	Corpora
BERT	**BookCorpus :** A broad base of general domain knowledge. **English Wikipedia:** Factual information and relationships between entities in diverse topics. **Parameter Size:** 110M
BioBERT	**PubMed:** Incorporates a vast collection of biomedical literature, including research articles and publications, from the PubMed database. **PMC:** Utilizes full-text articles from the PubMed Central (PMC) repository, enriching the model's understanding of biomedical language and concepts. **Parameter Size: 110M**
ClinicalBERT	**MIMIC-III**: Provides real-world clinical data, including notes and summaries from the MIMIC-III database, capturing diverse medical scenarios. **Parameter Size:** 110M
Bio+CLinicalBERT	**MIMIC-III**: Provides real-world clinical data, including notes and summaries from the MIMIC-III database, capturing diverse medical scenarios. **PubMed**: Incorporates biomedical literature from PubMed, offering a comprehensive understanding of medical research and terminology. **Parameter Size:** 110M

PubMedBERT	**PubMed**: Trained exclusively on PubMed abstracts and PubMed Central full-text articles, focusing on biomedical domain knowledge. **Parameter Size:** 110M
BioLinkBERT	**PubMed:** Incorporates biomedical literature from PubMed. **PMC:** Includes full-text articles from PubMed Central. Additionally may use other biomedical databases or resources to enhance linking between biomedical entities. **Parameter Size:** 110M

3.3 Proposed Framework

Our suggested framework for the biological extractive based factoid-type QA task is explained in this section. We begin by defining the issue and giving a solution by a general overview of the framework. We then go into further detail about the entire procedure.

3.3.1 Problem Scope and Solution Architecture

The biomedical factoid question-answering task is an extractive QA problem where the goal is to pinpoint the exact start and end positions of an answer within a context passage, given a specific question. Each context passage (C) has a length of m tokens, and the question (Q) consists of n tokens, with the objective being to extract a single, fact-based answer (A) from the passage. The challenge lies in handling complex biomedical language, where answers may be deeply embedded in dense, technical text, and require precise extraction of entities or relationships without deviating from the factual correctness. In addition to the inherent difficulty of parsing long and unstructured biomedical documents, the system must accurately interpret domain-specific terminologies, disambiguate between overlapping entities, and ensure that the answer identified corresponds precisely to the question. Given the high stakes in biomedical applications, the QA system must provide high accuracy and efficiency in determining the correct spans of text, minimizing any errors in the extraction process.

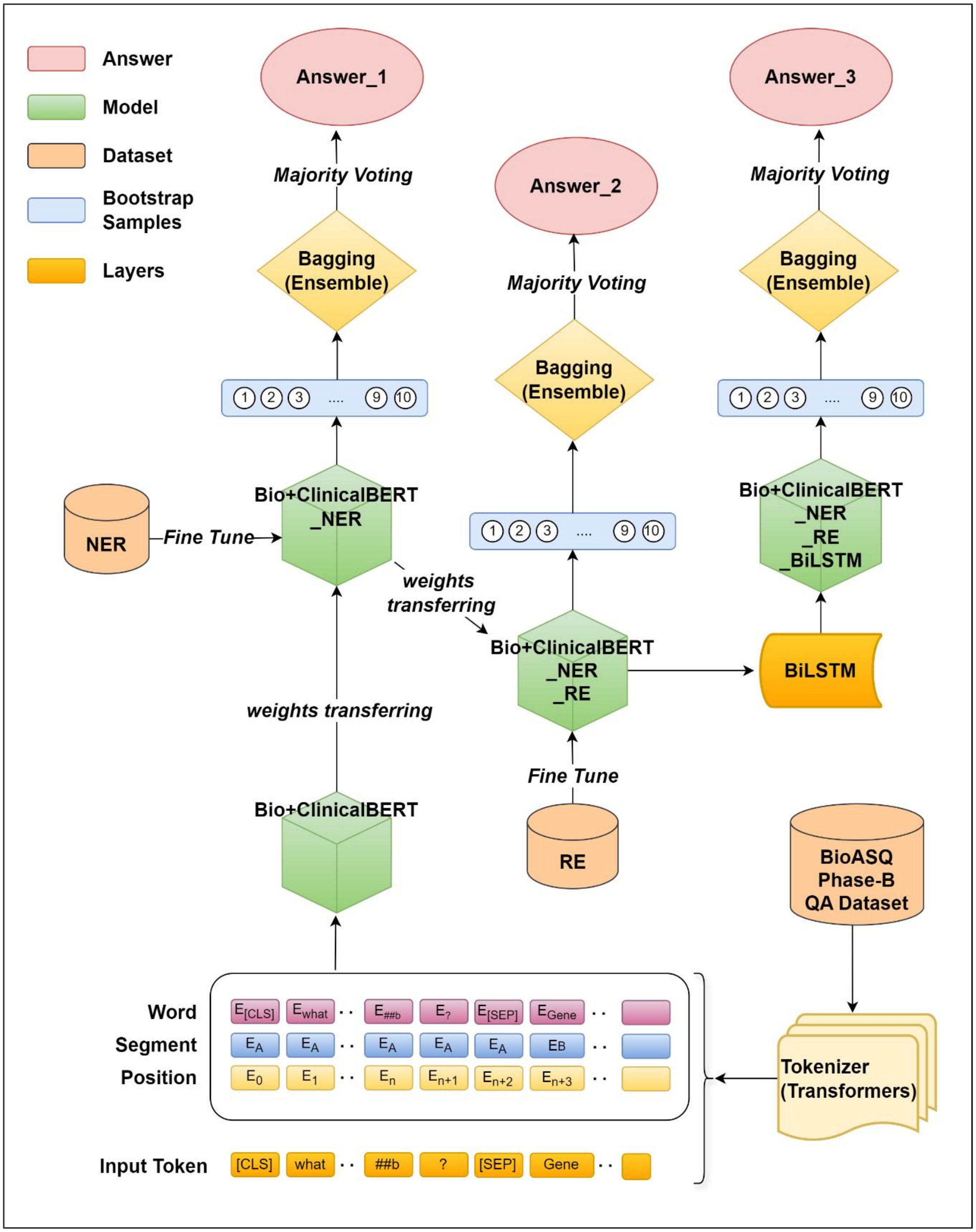


Figure 3. The proposed framework involves transferring weights from NER and RE models, incorporating a BiLSTM layer, and applying bootstrap sample aggregation for ensembling.

As shown in

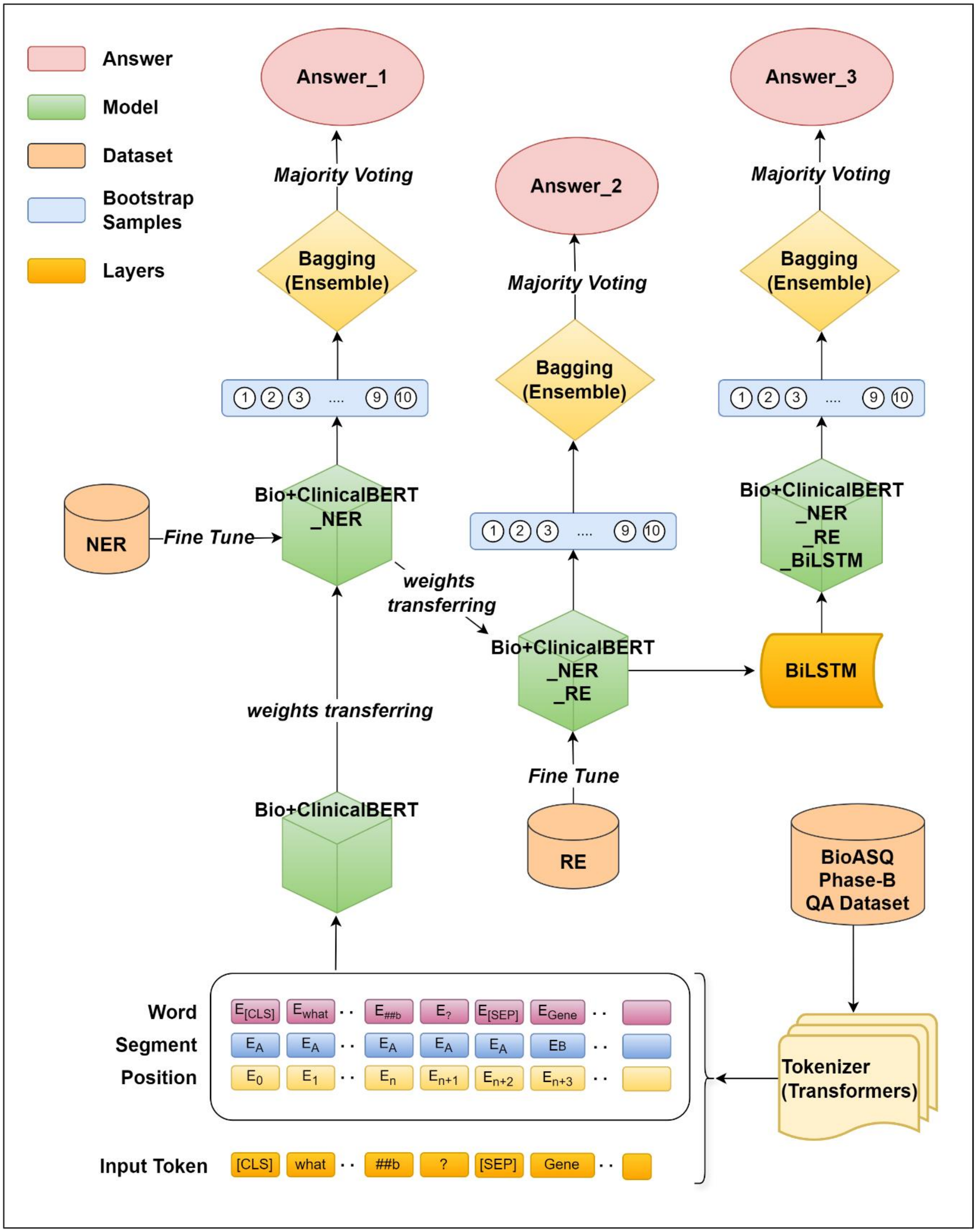


Figure 3 ,The proposed biomedical factoid question-answering framework integrates cutting-edge

NLP techniques in a multi-tiered, synergistic approach. At its core, Bio+ClinicalBERT provides domain-specific language understanding, enhanced through cascading transfer learning stages for Named Entity Recognition and Relation Extraction. A BiLSTM layer captures long-range dependencies, while ensemble learning via bagging mitigates overfitting. Moreover, the ensemble strategy, leveraging multiple instances of the Bio+ClinicalBERT_NER_RE_BiLSTM model trained on bootstrap samples, harnesses the collective wisdom of diverse model perspectives. This design not only extracts precise answers but also offers a framework for deeper analysis of the reasoning process, allowing for traceability and interpretability of the model's decisions. Additionally, it improves generalizability across varying biomedical contexts, enhancing the system's applicability to real-world clinical scenarios. Such an architecture not only meets the high accuracy demands of the biomedical field but also provides a valuable tool for researchers and clinicians seeking to extract actionable insights from complex medical data.

Table 11. Algorithm for our proposed BEQA system

```
Load Pre-trained Model:
Bio+ClinicalBERT= load_pretrained_model("Bio+ClinicalBERT")

NER and RE Model Fine-tuning:
Bio+ClinicalBERT_NER = finetune(Bio+ClinicalBERT, dataset="NER")
Bio+ClinicalBERT_RE = finetune(Bio+ClinicalBERT, dataset="RE")

QA Model Preparation:
Bio+ClinicalBERT_NER QA Finetuned= finetune(Bio+ClinicalBERT_NER, dataset="BioASQ_QA")
// Knowledge transferring from Bio+ClinicalBERT_NER to Bio+ClinicalBERT_RE
Bio+ClinicalBERT_NER_RE = transfer_knowledge(Bio+ClinicalBERT_NER, Bio+ClinicalBERT_RE)
Bio+ClinicalBERT_NER_RE QA Finetuned= finetune(Bio+ClinicalBERT_NER_RE, dataset="BioASQ_QA")
Bio+ClinicalBERT_NER_RE_BiLSTM = add_layer(Bio+ClinicalBERT_NER_RE, layer="BiLSTM")
Bio+ClinicalBERT_NER_RE_BiLSTM QA Finetuned = finetune(Bio+ClinicalBERT_NER_RE_BiLSTM, dataset="BioASQ_QA")

Ensemble Prediction:
function ENSEMBLE_PREDICT(input_text, question)
    predictions = []
    function bootstrap_sampling(data, n_samples)
        return [random.choices(data, k=len(data)) for _ in range(n_samples)]
    end function
    for model in [Bio+ClinicalBERT_NER, Bio+ClinicalBERT_NER_RE, Bio+ClinicalBERT_NER_RE_BiLSTM] do
        model_predictions = []
        bootstrap_samples = bootstrap_sampling(input_text, n_samples=10)
        for subset in bootstrap_samples do
            prediction = model.predict(subset, question)
            model_predictions.append(prediction)
        end for
        // Aggregate by Majority prediction
        aggregated_prediction = aggregate_predictions(model_predictions)
        predictions.append(aggregated_prediction)
    end for
    return predictions
end function
```

3.4 Fine-Tuning

3.4.1 Fine-tuning on QA Dataset

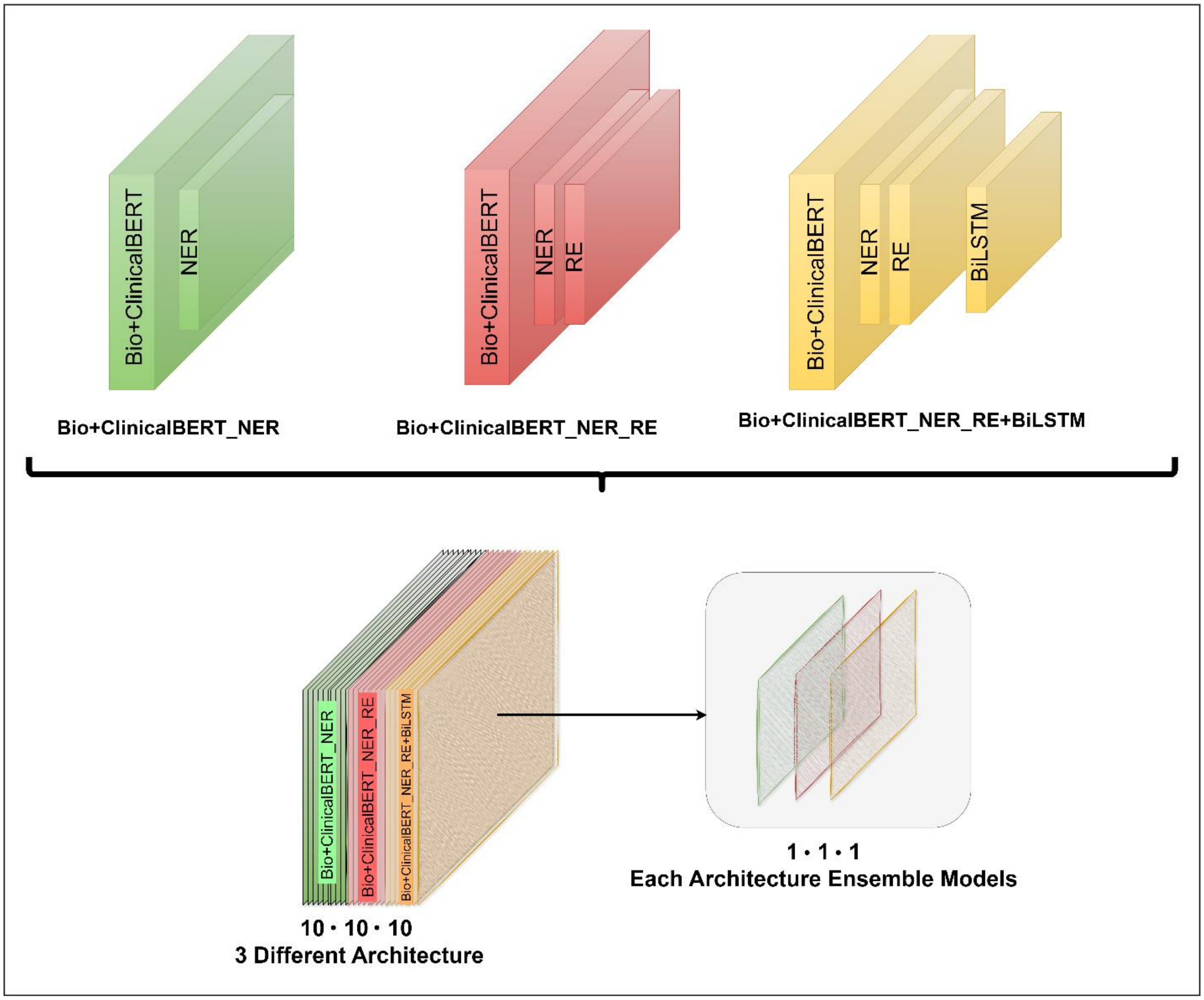
Extracting answers to questions from a given context is the task of question answering. We have fine-tuned six language representation models for our comparative analysis, which are BERT, BioBERT, Bio+Clinical BERT, ClinicalBERT, PubMedBERT, and BioLinkBERT. All models were fine-tuned using the uniform BERT architecture used for SQuAD formatted dataset. The BioASQ 4b factoid dataset was selected for this purpose due to its compatibility with the SQuAD format, streamlining the fine-tuning process for all six models that share a uniform architecture. This systematic approach not only ensured consistency in model architecture but also facilitated an in-depth evaluation across multiple dimensions, ultimately contributing to a robust and informed analysis. By comparing performance, we picked Bio+ClinicalBERT pre-trained model as our base work model.

3.4.2 Fine-tuned on NER and RE Dataset

Named Entity and Relation Extraction are critical for enhancing work performance. In order to improve Bio+ClinicalBERT's ability to learn about the named entities and determine their relationship, we employ NER and RE datasets by a method that have shown to be successful [37]. A novel approach utilizes Bio+ClinicalBERT models for NER and RE, merging their architectures and outputs. This Combined Model improves comprehension of complex relationships and captures long-range dependencies, enhancing accuracy in Question Answering tasks.

3.5 System Blueprint

The proposed Biomedical Extractive Question Answering (BEQA) system employs a multi-stage architecture that progressively enhances the base Bio+ClinicalBERT model through transfer learning, additional neural network layers, and ensemble techniques. The system blueprint in Figure 4 can be delineated into the following key components:



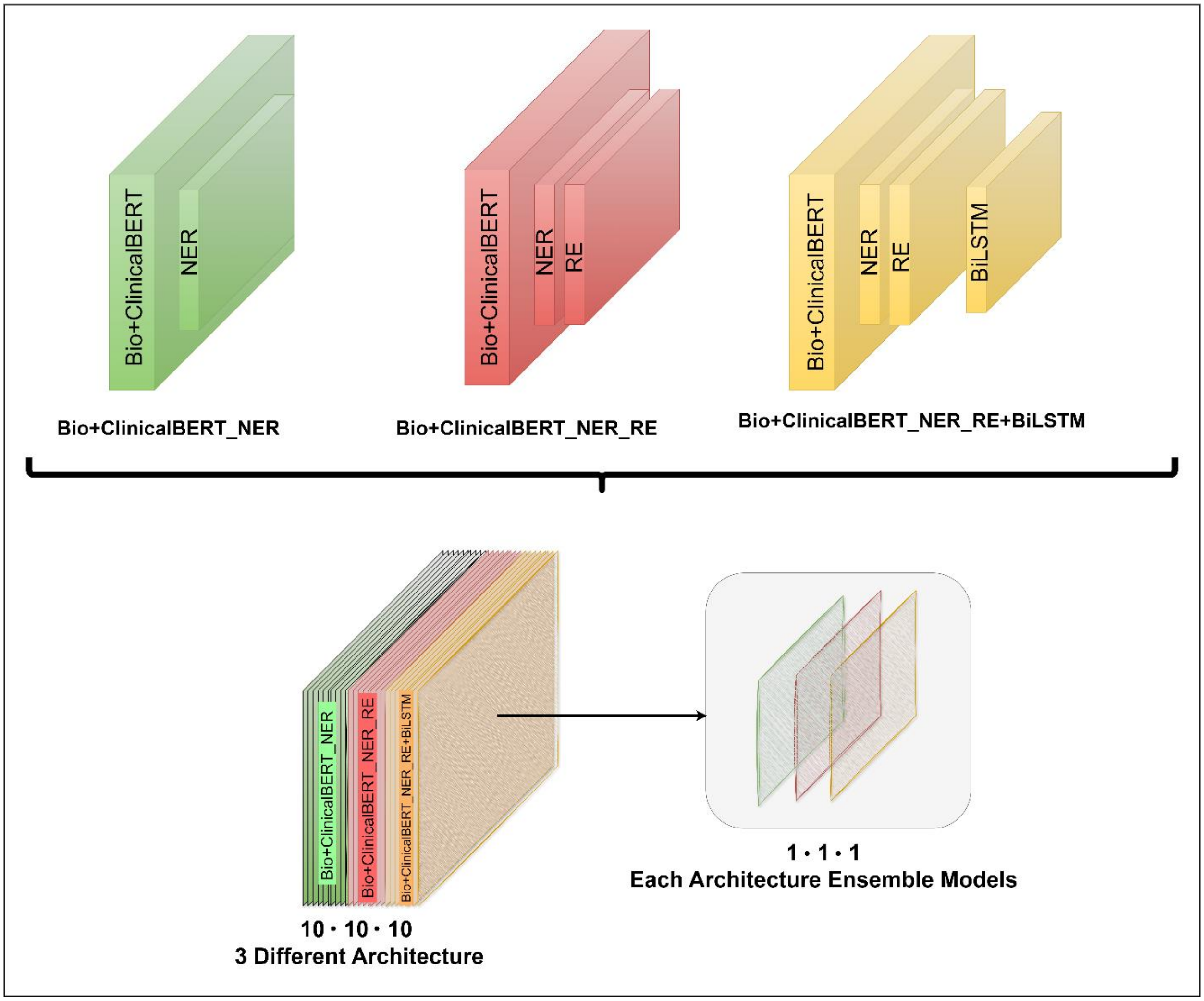


Figure 4. Multi-stage BEQA System Blueprint

The first stage, represented by the green block, shows Bio+ClinicalBERT_NER, where Named Entity Recognition (NER) capabilities are integrated into the base model. This is followed by the red block, Bio+ClinicalBERT_NER_RE, which further incorporates Relation Extraction (RE) on top of the NER-enhanced model. The final and most comprehensive model, shown in yellow, is Bio+ClinicalBERT_NER_RE+BiLSTM, which adds a Bidirectional Long Short-Term Memory (BiLSTM) layer to capture sequential dependencies in the biomedical text. These three architectural variants represent increasingly complex models designed to handle the nuances of biomedical language and extract precise answers.

The lower part of the figure depicts the ensemble learning strategy employed to boost performance and reliability. Each of the three model architectures is replicated 10 times, resulting in a 10 - 10 -

10 configuration (30 models in total). This approach leverages the power of ensemble learning, where multiple models collaborate to produce more accurate and robust predictions. The final stage, shown on the right, represents the 1 x 1 x 1 Each Architecture Ensemble Models, indicating that the outputs from the 30 models are aggregated to produce a single, highly refined answer for each architecture type. This ensemble approach helps mitigate individual model biases and improves the overall system's ability to handle complex biomedical questions and contexts.

The system design, featuring a hierarchical and modular structure, utilizes a transfer learning approach, multiple NLP tasks, and advanced neural architectures to tackle biomedical question answering complexities, resulting in a robust and high-performing BEQA solution.

3.5.1 Transfer Learning Influence

Transfer learning has become a cornerstone of modern machine learning, enabling models to leverage pre-existing knowledge from related tasks to improve performance on new challenges. Our framework employs a Task-Specific Sequential Transfer Learning approach, where knowledge is systematically transferred from simpler tasks like NER to more complex tasks such as RE and, ultimately, QA. This structured transfer of knowledge ensures that the model builds a robust understanding of biomedical text, layer by layer, making it highly effective at the final QA task.

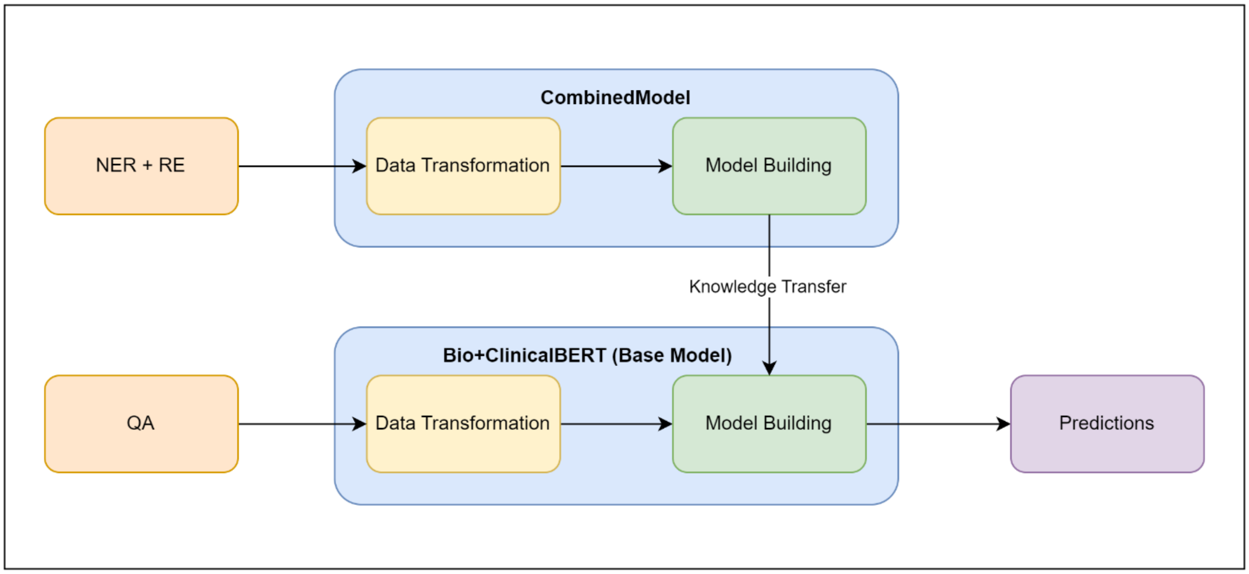


Figure 5. Transfer Learning Process

The process begins with fine-tuning a pre-trained Bio+ClinicalBERT model on the NER task, where it adjusts its weights and parameters to specialize in identifying biomedical entities. These learned weights and parameters, representing the acquired knowledge of entity recognition, are then transferred to the Relation Extraction (RE) task. The RE model starts with these optimized weights, providing a solid foundation in entity recognition, which is crucial for understanding and extracting relationships between entities. As the model is further fine-tuned on the RE dataset, the transferred weights are refined to capture relational knowledge. Subsequently, the learned weights and parameters from the RE task are again transferred to the Question Answering (QA) model. This sequential transfer ensures that the QA model does not start from scratch but leverages the cumulative knowledge of entity recognition and relation extraction, culminating in a QA model with a deep understanding of biomedical language and concepts. So, This three-stage transfer learning approach leverages knowledge from both tasks, resulting in a robust and accurate BEQA system.

3.5.2 Investigating BiLSTM Approach

In the context of BEQA models, incorporating a BiLSTM (Bidirectional Long Short-Term Memory) layer after completing transfer learning can significantly enhance the model’s performance. While BERT-based architectures excel in capturing rich contextual information from text, they may fall short in fully leveraging the sequential dependencies inherent in natural language. To address this, a BiLSTM layer is added on top of BERT’s output, enhancing the model's ability to understand the flow of information in both directions within a sequence. This approach is particularly beneficial for BEQA tasks, where accurately identifying answer spans often relies on understanding the relationships between entities and the order of information presented in the text.

To implement the BiLSTM layer, the architecture processes input text through BERT, generating contextualized embeddings (size 768). These embeddings are passed to a BiLSTM layer with a hidden size of 512 and 1 layer, which processes the data bidirectionally, resulting in a 2*512 output. This output is fed into a linear layer that produces logits with a dimension of 2, representing the start and end probabilities of an answer span. This setup enhances the model's ability to capture sequential dependencies for accurate answer extraction in BEQA tasks.

3.5.3 Ensemble Learning for Enhanced System

In the context of BEQA, where precision and robustness are paramount, ensemble learning offers a promising approach to enhance model performance. Ensemble learning involves combining multiple models to produce a single, more accurate predictive model. This technique is particularly useful in BEQA systems, where slight variations in model predictions can lead to significantly different outcomes. Among the various ensemble methods, Bagging (Bootstrap Aggregating) is a notable technique that reduces variance and improves generalization, which is critical in biomedical applications. By integrating Bagging with majority voting, the proposed BEQA system leverages the strengths of multiple models to deliver more reliable and robust answers to complex biomedical questions. In our proposed Biomedical Extractive Question Answering (BEQA) system, we leverage ensemble learning techniques to further enhance the performance and robustness of our models. Specifically, we employ bootstrap aggregation (bagging) combined with majority voting to create a powerful ensemble of the three main BEQA architectural models individually: Bio+ClinicalBERT_NER, Bio+ClinicalBERT_NER_RE, and Bio+ClinicalBERT_NER_RE_BiLSTM.

The ensemble learning process begins with bootstrap sampling, a technique that creates multiple subsets of our training data by randomly sampling with replacement. Following ten bootstrap samplings, we proceed with model training. Each of the 30 models (10 for each model architecture) is used to train an independent instance of its respective BEQA model. This results in 10 variants of Bio+ClinicalBERT_NER, 10 variants of Bio+ClinicalBERT_NER_RE, and 10 variants of Bio+ClinicalBERT_NER_RE_BiLSTM. The independent training on slightly different datasets ensures that each model instance develops unique characteristics and potentially captures different aspects of the underlying data distribution. The final stage of our ensemble learning approach involves aggregating the predictions from all 10 model instances for each model architecture using majority voting. When presented with a new question and context, each model in the ensemble generates its prediction for the answer span. The system then determines the final answer by selecting the span that receives the most votes across all models.

This ensemble architecture leverages the strengths of multiple model variants and training instances, resulting in a more robust and accurate BEQA system. The diversity introduced by different architectures (NER, NER+RE, NER+RE+BiLSTM) and bootstrap samples helps to mitigate individual model biases and errors, leading to improved generalization and performance on biomedical question answering tasks.

Chapter 4 Investigation/Experiment, Result, Analysis and Discussion

4.1 Evaluation Metrics

In our research, we employed several key assessment metrics to fully evaluate our model's performance across various natural language processing tasks. Each metric provides unique insights into the model's capabilities and effectiveness in handling specific challenges inherent in the tasks at hand. To evaluate the performance of our biomedical extractive question-answering system, we employed the following metrics: Exact Match, F1 Score and Lenient Accuracy.

4.1.1 Exact Match (EM)

This metric measures the percentage of answers that exactly match the ground truth answer. EM is a binary score, where 1 if the predicted answer is identical to the ground truth, 0 otherwise.

- EM = (Number of Exact Matches) / (Total Number of Questions)

EM is crucial for applications where precision is paramount, as it reflects the model's ability to provide exactly correct answers. Ensures that the model provides precise and accurate biomedical information, which is essential in medical contexts where incorrect answers can have serious implications.

4.1.2 F1 Score

F1-score is the harmonic mean of precision and recall. It considers both the exactness and completeness of the answer. To calculate F1-score, the predicted and ground truth answers are typically tokenized.

- Precision = (Number of Correct Words) / (Number of Predicted Words)
- Recall = (Number of Correct Words) / (Number of Ground Truth Words)
- F1 = 2 * (Precision * Recall) / (Precision + Recall)

The F1 score is significant for imbalanced datasets where precision and recall may not be equally weighted. Balances the importance of retrieving all relevant biomedical answers (recall) while minimizing the retrieval of incorrect ones (precision), which is critical for medical accuracy.

4.1.3 Lenient Accuracy (LAcc)

LACC measures the overlap between the predicted answer and the ground truth answer. It calculates the length of the longest common subsequence (LCS) between the two sequences and divides it by the length of the ground truth answer.

- LACC = LCS (Predicted Answer, Ground Truth Answer) / Ground Truth Answer

This metric is significant because it allows partial credit for including the accurate response from the list, even if it is not the top-ranked answer. LAcc is important for ensuring the model can include the correct answer among its predictions, reflecting its ability to identify relevant information even if the exact placement is not prioritized.

By leveraging these diverse evaluation metrics, we gain a comprehensive understanding of our model's performance across different dimensions, enabling us to draw meaningful insights and conclusions regarding its efficacy in addressing the specific challenges posed by natural language processing tasks.

4.2 Experimental Configuration

This study tests a proposed framework on BioASQ datasets, which include 327 factoid questions and 3266 question passage pairs. The BioASQ 4b and 9b training datasets were pre-processed, and several pre-trained Base models were used as baselines. The model with the most stable and superior performance was chosen for further enhancement, Bio+ClinicalBERT.

We additionally utilized two different categories of datasets to fine-tune Bio+ClinicalBERT. Firstly, we use the NER datasets, which is a collection fully annotated at the mention and concept levels. Secondly In the RE datasets, which is collection of annotated labeling data of relation between entities. We implemented the experiments on a system equipped with NVIDIA GPUs GTX-1070, and fine-tuning was performed using the Adam optimizer with :

- Batch Size: 8, 16
- Learning Rates: 2e5
- Epoch: 10
- Loss Function: CrossEntropy Loss
- N_BEST = 20

- MAX_Answer_Length = 30

A machine learning model's key factors include batch size, learning rate, epochs, CrossEntropy Loss, n_best, and max_answer_length. Batch sizes range from 8 to 16, impacting memory usage and convergence speed. A learning rate of 2e-5 promotes stable model weight updates. The number of epochs determines the number of times the model runs through the dataset. CrossEntropy Loss guides the model to predict probabilities closer to actual labels. The N_BEST hyperparameter sets the number of possible answers, while the MAX_Answer_Length parameter sets the maximum length for extracted answers, ensuring concise and relevant predictions.

4.3 Model Comparison Interpretation

4.3.1 Baseline Methods

In this study, We employed several pre-trained models as our baseline methods, including BERT, BioBERT, PubMedBERT, BioLinkBERT, ClinicalBERT, and Bio+ClinicalBERT. in

Table 12. Performance of the BERT models on the BioASQ 4b Factoid QA dataset. The Bio+ClinicalBERT model achieved the highest performance, with an F1 score of 86.24 and an Exact Match (EM) score of 84.09 after 10 epochs, outperforming all other models on this dataset. summarize the performance of all baselines on the BioASQ 4b dataset. By looking the table, we observe:

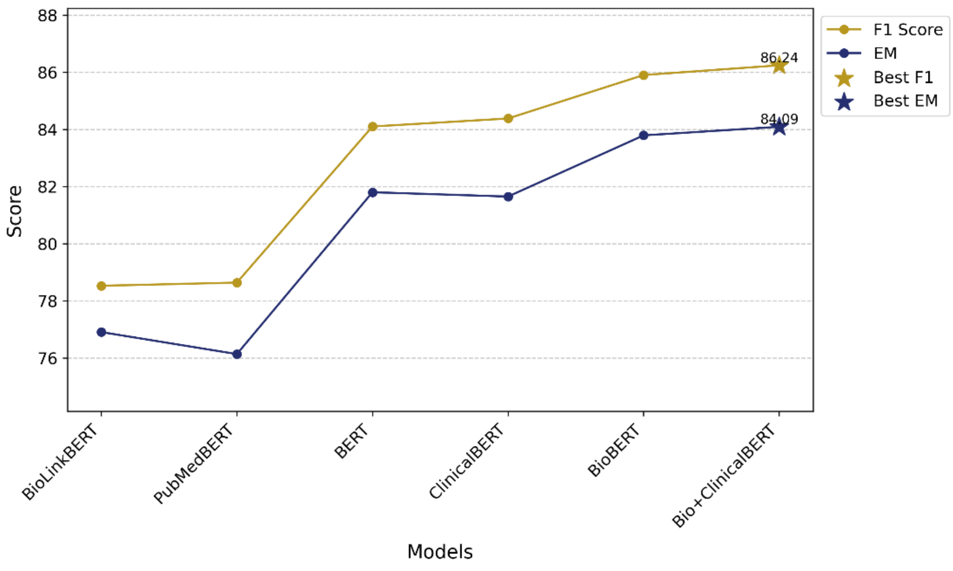


Figure 6. Comparison of Baseline Models by F1 and EM

Table 12. Performance of the BERT models on the BioASQ 4b Factoid QA dataset. The Bio+ClinicalBERT model achieved the highest performance, with an F1 score of 86.24 and an Exact Match (EM) score of 84.09 after 10 epochs, outperforming all other models on this dataset.

Dataset: BioASQ 4b (factoid)			
Model	**Epoch**	**F1 Score**	**EM**
BERT	10	84.10	81.80
BioBERT	7	85.90	83.79
PubMedBERT	3	78.64	76.14
BioLinkBERT	5	78.53	76.91
ClinicalBERT	8	84.38	81.65
Bio+ClinicalBERT	10	**86.24**	**84.09**

The table presents the performance of various BERT-based models on the BioASQ 4b factoid QA dataset. The Bio+ClinicalBERT model is the highest-performing, achieving an F1 score of 86.24 and an EM score of 84.09 after 10 epochs. This is due to its ability to leverage biomedical and clinical textual data, enhancing its understanding of domain-specific terminology and contextual nuances. The robustness of Bio+ClinicalBERT makes it the most stable and effective choice for further experimentation in the BEQA system, aligning with the overall methodology of selecting and fine-tuning models for high performance on domain-specific tasks. The comparison underscores the importance of selecting an appropriate base model for biomedical QA tasks and justify our choice of Bio+ClinicalBERT as the foundation for our advanced BEQA system, which we further enhanced through transfer learning, BiLSTM integration, and ensemble techniques to push the boundaries of performance in biomedical question answering.

4.3.2 Ablation Study

To evaluate the relative contributions of the different parts of our proposed framework, we performed an ablation study by Transfer Learning and Adding BiLSTM layer techniques. The below table and figure shows the ablation study that compares three design alternatives.

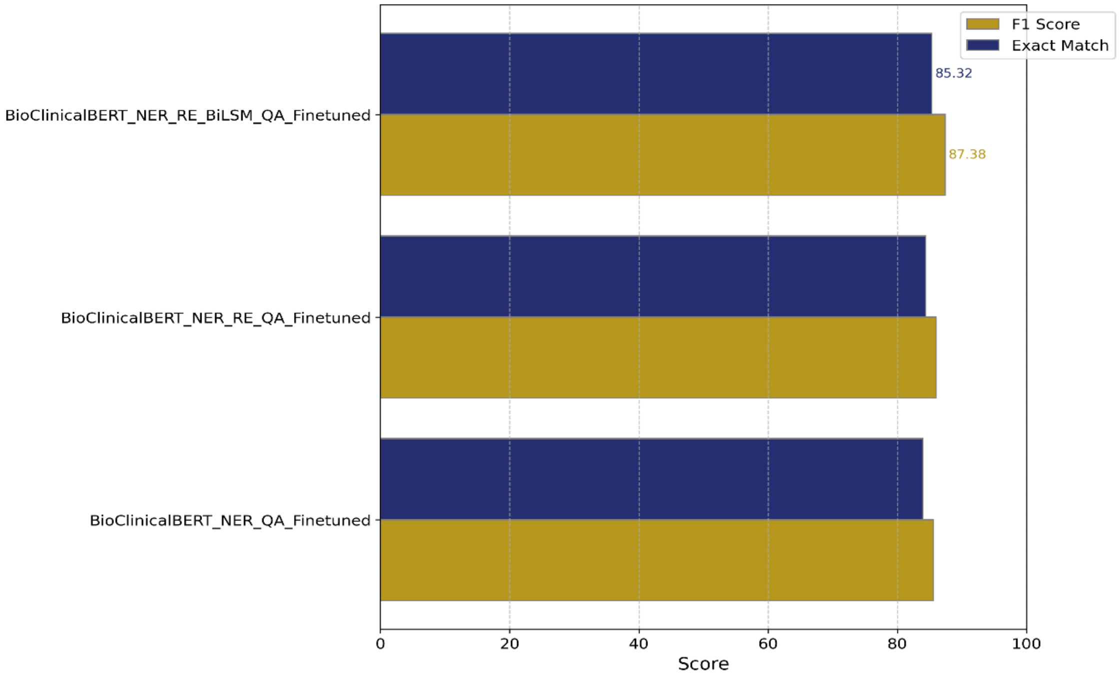


Figure 7. Comparing Performance of Transfer Learning and Adding BiLSTM Layer techniques in F1 & EM score on BioASQ 4b Factoid QA dataset

Table 13. Ablation Study on the BioASQ 4b Factoid QA dataset. The Bio+ClinicalBERT_NER_RE_BiLSTM model achieved the best performance, with an average F1 score of 84.88 and an EM score of 82.78. At epoch 2, it reached the highest scores, with an F1 of 87.38 and and an EM of 85.32. The other model variants, Bio+ClinicalBERT_NER and Bio+ClinicalBERT_NER_RE, showed slightly lower performance across the study.

Epoch	**Bio+ClinicalBERT_NER QA Finetuned**		**Bio+ClinicalBERT_NER_RE QA Finetuned**		**Bio+ClinicalBERT_ NER_RE_BiLSM QA Finetuned**	
	F1	EM	F1	EM	F1	EM
1	72.66	60.85	78.48	75.69	82.49	78.44
2	80.71	77.52	**86.02**	**84.40**	**87.38**	**85.32**
3	83.87	81.19	79.99	78.29	83.20	81.19

4	84.74	81.65	81.96	79.05	84.53	82.11
5	83.31	79.97	85.02	82.87	86.40	85.02
6	84.09	81.80	82.86	80.43	85.93	83.94
7	83.19	81.19	81.43	79.20	84.63	82.11
8	84.75	83.03	84.48	82.26	82.47	80.73
9	85.08	83.49	83.77	81.50	85.86	84.40
10	**85.58**	**83.94**	83.76	81.35	85.88	84.56
AVG	**82.80**	**79.46**	**82.78**	**80.50**	**84.88**	**82.78**

The ablation study presented in the table demonstrates the impact of different architectural choices within the proposed BEQA system, specifically focusing on the sequential transfer learning and the addition of a BiLSTM layer. The study compares three design alternatives: Bio+ClinicalBERT_NER QA Finetuned, Bio+ClinicalBERT_NER_RE QA Finetuned, and Bio+ClinicalBERT_NER_RE_BiLSTM QA Finetuned. The results reveal that the model incorporating both NER and RE tasks (Bio+ClinicalBERT_NER_RE QA Finetuned) already shows improvement over the model trained solely on NER, highlighting the benefits of transferring knowledge across tasks. However, the introduction of the BiLSTM layer in the Bio+ClinicalBERT_NER_RE_BiLSTM model further enhances performance, with an average F1 score of 84.88 and an EM score of 82.78 over 10 epochs. Notably, at epoch 2, this model achieves its highest F1 and EM scores of 87.38 and 85.32, respectively.

This ablation study validates our approach, confirming that the combination of transfer learning techniques and advanced neural architectures leads to superior performance in biomedical extractive question answering tasks. The BiLSTM-based model (Bio+ClinicalBERT_NER_RE_BiLSTM) consistently outperforms other configurations, showcasing its robustness in handling complex biomedical queries.

4.4 Model Performance Insights

4.1 Quantitative Results

We now compare the suggested framework to the ablation study models in ensembled bagging approach. Here we used 9b datasets for better performance because of larger size data than 4b. The result of each model is summarized in in Table 14.

Table 14. BEQA models are (A) Bio+ClinicalBERT_NER QA Finetuned , (B) Bio+ClinicalBERT_NER_RE QA Finetuned, and (C) Bio+ClinicalBERT_NER_RE_BiLSTM QA Finetuned. Performance Comparison of Proposed Framework Models on the BioASQ 9b Factoid QA Dataset. This table shows the F1, Exact Match (EM), and Lenient Accuracy (LAcc) scores for three Bio+ClinicalBERT model variants fine-tuned on the QA task across 10 bootstrap samples. The Bio+ClinicalBERT_NER_RE_BiLSTM model achieved the highest ensemble scores with an F1 of 91.55, an EM of 88.62, and an LAcc of 0.84, outperforming the other two model variants, Bio+ClinicalBERT_NER and Bio+ClinicalBERT_NER_RE, in overall accuracy and consistency across the experiments.

Experiments		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Ensemble Score**
A	**F1**	90.42	90.05	90.20	**90.91**	90.06	90.35	90.47	89.69	90.00	89.40	89.89
	EM	86.51	86.33	87.43	87.16	86.24	87.16	86.51	86.51	**87.52**	85.78	85.87
	LAcc	0.82	0.81	0.83	**0.83**	0.82	0.83	0.82	0.82	0.83	0.81	0.81
B	**F1**	89.89	89.94	**91.15**	91.07	90.91	90.23	90.53	90.86	90.45	90.18	90.01
	EM	86.05	86.06	**88.17**	87.89	87.43	86.79	86.70	87.71	87.16	87.16	85.96
	LAcc	0.82	0.82	**0.84**	0.84	0.83	0.82	0.82	0.83	0.82	0.82	0.81
C	**F1**	90.35	90.67	91.04	90.98	**91.69**	90.85	90.76	90.37	90.23	90.07	**91.55**
	EM	86.88	87.25	**88.35**	87.16	87.80	87.43	87.06	86.88	86.97	86.79	**88.62**
	LAcc	0.83	0.82	**0.84**	0.83	0.83	0.83	0.83	0.82	0.82	0.82	**0.84**

4.2 Qualitative Analysis

The analysis of Table 14 showcases the performance comparison of three BEQA models (A) Bio+Clinical BERT_NER, (B) Bio+ClinicalBERT_NER_RE, and (C) Bio+ClinicalBERT_NER_RE_BiLSTM—across 10 bootstrap samples, with each model evaluated on F1, Exact Match (EM), and Lenient Accuracy (LAcc). Additionally, ensemble scores are provided, representing the overall performance of each model.

Across the 10 samples, model C (Bio+ClinicalBERT_NER_RE_BiLSTM) achieves the highest F1 score of 91.55 in the ensemble, outperforming model B at 90.01 and model A at 89.89. Within the individual experiments, model C leads with the highest F1 score 91.69 in experiment 5. While model A peaks at 90.91 in Experiment 4, and model B reaches 91.15 in Experiment 3 as well. Although model B shows a competitive F1 score in Experiment 3, model C maintains an overall advantage in most samples. In case of EM metric, Model C excels once more, scoring an ensemble score of 88.62, which is greater than the 85.96 and 85.87 of models B and A. Model C achieves the highest EM of 88.35 out of the 10 samples. Model A and model B peak at 87.52 (Experiment 9) and 88.17 (Experiment 3), respectively. Also for LAcc, model C performs better than models A and B, which both have ensemble scores of 0.81, with a score of 0.84. Conversely, in experiment 3, Models B and C tie for the highest individual LAcc score of 0.84, and in experiment 4, Model places 0.83. This table demonstrates clearly that model C is the best overall performance, consistently achieving higher scores across most experiments and maintaining the highest ensemble scores in all metrics. This detailed comparison further emphasizes the significance of model C’s architecture and its superior capabilities in handling the BioASQ 9b Factoid QA task.

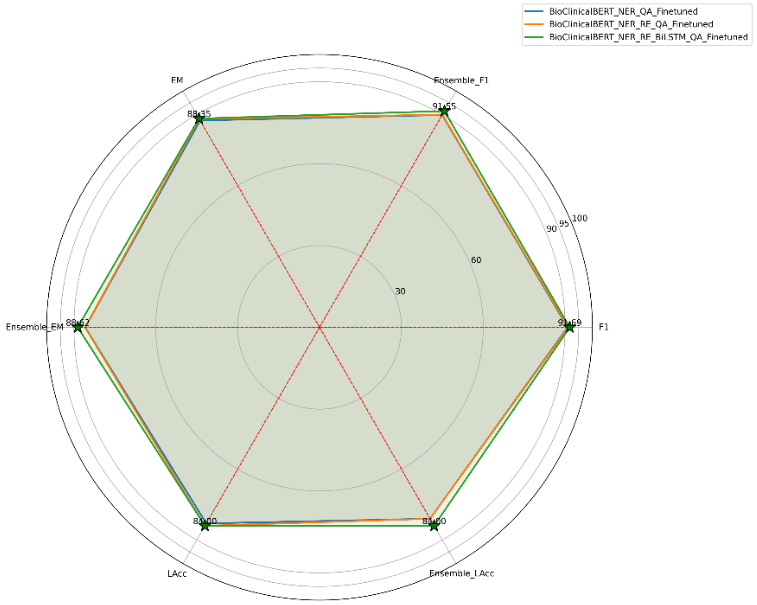


Figure 8. Comparing Performance of Ensembled of Transfer learning of NER and RE and BiLSTM layer adding frameworks in F1, EM, and LAcc score on BioASQ 9b Factoid QA dataset

A closer examination reveals that while the addition of Relation Extraction to NER yielded modest gains, the integration of BiLSTM resulted in the most significant performance boost. The BiLSTM model also exhibited superior consistency across all experiments, highlighting its enhanced stability. Quantitatively, the progression from NER solo to NER+RE+BiLSTM resulted in substantial improvements: a 1.66 point increase in F1 score, a 2.75 point rise in EM score, and a 0.03 point uplift in LAcc. The radar graph in Figure 8 visually corroborates these findings, with the BiLSTM model's polygon encompassing a notably larger area across all metrics. While the BiLSTM model demonstrates overall superiority, the marginal gains in LAcc suggest that the primary benefits are most pronounced in F1 and EM scores, offering valuable insights for future model refinements in biomedical question answering tasks. The framework, which combines transfer learning, advanced neural architectures, and ensemble techniques, significantly enhances the capability of biomedical extractive question answering systems. The incremental improvements from NER to NER+RE to NER+RE+BiLSTM highlight the cumulative benefits of each component. The ensemble scores for all models surpass their best individual experiment scores, proving the bagging approach's effectiveness in enhancing model robustness and performance.

4.5 Case Study and User Interface Evaluation

4.5.1 Case Study Exploration

This study presents a test case demonstrating the effectiveness of adjusting Bio+ClinicalBERT using knowledge transfer from Named Entity Recognition and Relation Extraction models, integrating a BiLSTM layer, and employing ensemble learning.

Question		**Which domain allowing self-association do exist in TDP-43 and FUS proteins?**
Context		Multistep process of FUS aggregation in the cell cytoplasm involves RNA-dependent and RNA-independent mechanisms. Fused in sarcoma (FUS) is an RNA-binding protein involved in pathogenesis of several neurodegenerative diseases. Aggregation of mislocalized FUS into non-amyloid inclusions is believed to be pivotal in the development of cell dysfunction, but the mechanism of their formation is unclear. Using transient expression of a panel of deletion and chimeric FUS variants in various cultured cells, we demonstrated that FUS accumulating in the cytoplasm nucleates a novel type of RNA granules, FUS granules (FGs), that are structurally similar but not identical to physiological RNA transport granules. **Formation of FGs requires FUS N-terminal prion-like domain and the ability to bind specific RNAs.** Clustering of FGs coupled with further recruitment of RNA and proteins produce larger structures, FUS aggregates (FAs), that resemble but are clearly distinct from stress granules. In conditions of attenuated transcription, FAs lose RNA and dissociate into RNA-free FUS complexes that become precursors of large aggresome-like structures. We propose a model of multistep FUS aggregation involving RNA-dependent and RNA-independent stages. This model can be extrapolated to formation of pathological inclusions in human FUSopathies.
Ground Truth Answer		**prion-like domain**
Bio+ClinicalBERT QA Finetuned	**Predicted Answer:** -terminal prion-like domain (Confidence: 0.00)	**Comment:** For the given question on protein self-association domains, the baseline Bio+ClinicalBERT model, returned an answer of "-terminal prion-like domain" with a confidence score of 0.00. However, the correct answer, identified as "prion-like domain," was not recognized adequately, highlighting the limitations of the baseline model in capturing specific biomedical terminology.

Bio+ClinicalBERT_NER QA Finetuned Model	**Predicted Answer:** N-terminal prion-like (Confidence: 0.14)	**Comment:** Upon fine-tuning with a NER dataset specialized in biomedical entities, our Bio+ClinicalBERT_NER model improved its performance. It returned the answer “N-terminal prion-like “, which closely aligns with the golden answer, achieving a confidence score of 0.14. This enhancement demonstrates the effectiveness of integrating NER knowledge to better recognize domain-specific terms in biomedical contexts.
Bio+ClinicalBERT_NER_RE QA Finetuned Model	**Predicted Answer:** prion-like domain (Confidence: 0.14)	**Comment:** To further enhance answer extraction, we expanded our model to Bio+ClinicalBERT_NER_RE QA Finetuned, incorporating knowledge from both NER and RE tasks. This model returned the answer "prion-like domain," with a confidence score of 0.14. Although the correct answer was retrieved, the confidence score indicated room for improvement in precision and certainty.
Bio+ClinicalBERT_NER_RE_BiLSTM QA Finetuned	**Predicted Answer:** prion-like domain (Confidence: 0.91)	**Comment:** To address the confidence score issue and enhance model precision, we introduced a BiLSTM layer to encode question text and capture deeper semantic relationships. This enhancement significantly boosted our model's performance, achieving an answer of "prion-like domain “ with a confidence score of 0.91. This result underscores how well our method performed in outperforming the capabilities of the baseline model and accurately extracting complex biomedical named entities.

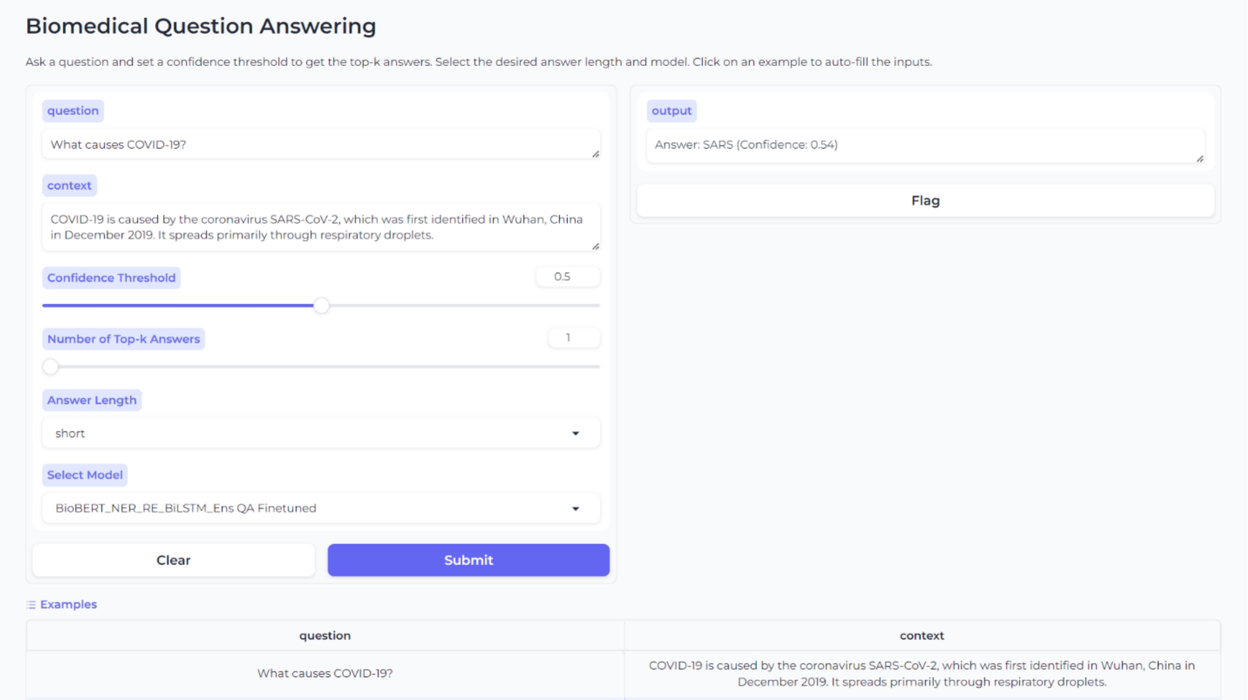
Table 15. Test case for Case Study

The case study presented in Table 15. Test case for Case Study highlights the performance of four BEQA models on a biomedical factoid question regarding the domain responsible for self-association in TDP-43 and FUS proteins. The models are tasked with identifying the correct domain based on a provided context, with the ground truth answer being "prion-like domain." The Bio+ClinicalBERT QA Finetuned model struggled with a question involving a -terminal prion-like domain, indicating its difficulty in understanding specialized biomedical terminology. The second model, Bio+ClinicalBERT_NER QA Finetuned, incorporated Named Entity Recognition (NER) and predicted "N-terminal prion-like" with a slightly higher confidence score. The third model, Bio+ClinicalBERT_NER_RE QA Finetuned, also included Relation Extraction, but produced the correct answer with the same confidence score. The final model, Bio+ClinicalBERT_NER_RE_BiLSTM QA Finetuned, provided the correct answer with a significantly higher confidence score of 0.91, indicating the BiLSTM layer's enhanced precision and overall accuracy, outperforming the previous models.

In summary, the introduction of NER, RE, and the BiLSTM layer led to gradual improvements in answer accuracy and confidence, illustrating the effectiveness of these layers in handling complex biomedical question answering tasks.

4.5.2 Interface Interaction Pipeline

Biomedical question-answering systems are essential for retrieving concise and accurate information from large biomedical datasets. The depicted interface enables users to query a biomedical model, providing a mechanism for customized answer generation based on specific inputs and configurable settings. Each option in this interface serves a specific function, contributing to the overall utility and flexibility of the system. Users can enter their questions and relevant context into specific fields. They can also adjust settings like confidence level, the number of top answers, answer length, and choose which model to use. After setting everything up, users can submit their question or reset the fields with a clear button. The system will display the answers in the output section, where users can also flag any incorrect answers. An example section provides sample questions and contexts to help users understand how the system works.



Biomedical Question Answering

Ask a question and set a confidence threshold to get the top-k answers. Select the desired answer length and model. Click on an example to auto-fill the inputs.

question: What causes COVID-19?

context: COVID-19 is caused by the coronavirus SARS-CoV-2, which was first identified in Wuhan, China in December 2019. It spreads primarily through respiratory droplets.

Confidence Threshold: 0.5

Number of Top-k Answers: 1

Answer Length: short

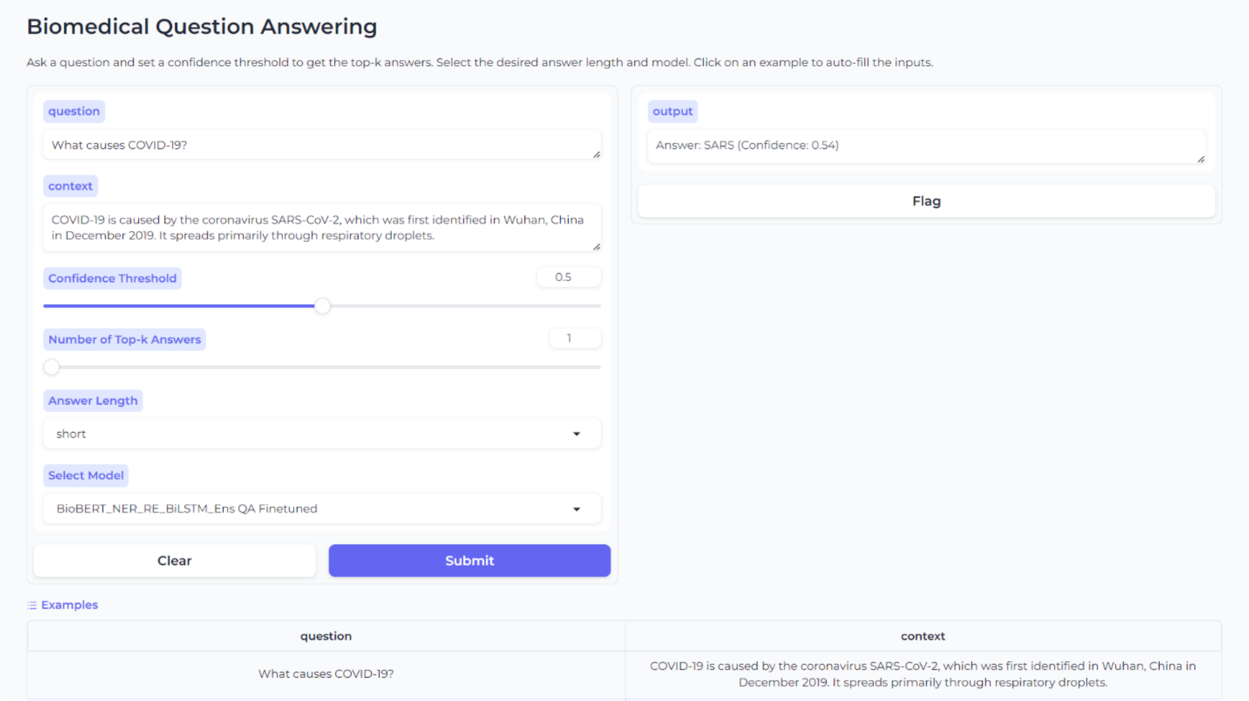
Select Model: BioBERT_NER_RE_BiLSTM_Ens QA Finetuned

output: Answer: SARS (Confidence: 0.54)

Examples

question	context
What causes COVID-19?	COVID-19 is caused by the coronavirus SARS-CoV-2, which was first identified in Wuhan, China in December 2019. It spreads primarily through respiratory droplets.

Figure 9. User Interface of BEQA



Biomedical Question Answering

Ask a question and set a confidence threshold to get the top-k answers. Select the desired answer length and model. Click on an example to auto-fill the inputs.

question: What causes COVID-19?

context: COVID-19 is caused by the coronavirus SARS-CoV-2, which was first identified in Wuhan, China in December 2019. It spreads primarily through respiratory droplets.

Confidence Threshold: 0.5

Number of Top-k Answers: 1

Answer Length: short

Select Model: BioBERT_NER_RE_BiLSTM_Ens QA Finetuned

output: Answer: SARS (Confidence: 0.54)

Examples

question	context
What causes COVID-19?	COVID-19 is caused by the coronavirus SARS-CoV-2, which was first identified in Wuhan, China in December 2019. It spreads primarily through respiratory droplets.

The interface shown in Figure 9 is for a Biomedical Question Answering system where users interact with an AI model to get answers based on biomedical queries. The pipeline begins with an input section where users can ask questions and optionally provide context to enhance the model's understanding, as seen with the question "What causes COVID-19?" and the associated description of the virus. Users can adjust the Confidence Threshold to set the minimum confidence level for the model's answer. There is also a slider to limit the number of top-k answers the model provides. The user can select the length of the answer, such as "short," and choose from available models, like BioBERT_NER_RE_BiLSTM QA Finetuned (bagging model). Once the configuration is done, the user submits the query, and the system returns an output with a confidence score, as seen with the output "SARS (Confidence: 0.54)." This interface allows fine-tuning of the AI's behavior based on user preferences.

The biomedical question-answering interface depicted in the image is structured to allow users significant control over the query process, facilitating accurate and context-aware answers. Through customizable settings such as confidence thresholds, top-k answers, and answer length, users can tailor their experience to meet specific research or clinical needs. This type of system could play an important role in automating the retrieval of information from biomedical datasets, improving both speed and accuracy in answering domain-specific queries.

Chapter 5 Impacts of the Project

5.1 Impact of this project on societal, health, safety, legal and cultural issues

Our research project on BioMedical Extractive Question Answering aims to improve how we find and understand information in biomedical research and healthcare. By combining different techniques like Named Entity Recognition (NER), Relation Extraction (RE), BiLSTM layers, and ensemble learning, our model tries to make it easier and more accurate to find important information in large biomedical datasets. This could have a big impact on society. It could speed up how quickly we learn new things in biomedical research by helping researchers find and understand information in scientific papers and medical records faster. This could mean that new treatments, medicines, and medical procedures are developed more quickly, which could improve how well patients do and even save lives. It could also help doctors and other healthcare workers make better decisions about diagnosing diseases and predicting how well treatments will work for individual patients, improving overall healthcare. Plus, by encouraging collaboration and sharing knowledge among scientists and healthcare workers, our project could lead to even more breakthroughs in solving big health problems. Overall, our research project could make healthcare more accurate, efficient, and accessible for everyone.

5.2 Impact of this project on environment and sustainability

Our project on BioMedical Extractive Question Answering has a strong potential for a viable business model, feasibility, and financial scalability. Here's how we envision it

Educational Licensing: We can offer discounted or free access to our platform for medical colleges and universities, allowing students to utilize it as part of their coursework and research projects. This model not only benefits students by providing valuable learning resources but also helps in establishing our technology as a standard tool within medical education programs.

Integration with Medical Curriculum: Collaborating with medical educators to integrate our platform into existing coursework can ensure seamless adoption and relevance within the academic curriculum. By aligning with learning objectives and competencies, we can demonstrate the utility of our technology in augmenting traditional teaching methods and fostering critical thinking skills among students.

Service Offering: We can offer our advanced biomedical text analysis as a service to research institutions, pharmaceutical companies, and healthcare organizations. These entities often struggle with managing and extracting insights from vast amounts of biomedical data. Our solution can help them streamline their research processes, accelerate drug discovery, and improve patient care through more informed decision-making.

Subscription Model: Implementing a subscription-based pricing model would provide our clients with access to our platform on a recurring basis. This model ensures a steady stream of revenue while offering flexibility for clients to scale their usage according to their needs.

Custom Solutions: We can offer tailored solutions and consulting services to clients with specific needs or datasets. This could include customized models trained on proprietary data or specialized analysis pipelines designed to extract domain-specific insights.

Partnerships and Collaborations: Collaborating with academic institutions, research labs, and industry partners can help expand our reach and access to diverse datasets. Partnerships with healthcare providers can also facilitate the integration of our technology into clinical workflows, opening up new revenue streams and opportunities for impact.

Scalability: As demand for biomedical data analysis continues to grow, we can scale our operations by investing in infrastructure, expanding our team of data scientists and domain experts, and continuously improving our technology through research and development efforts.

Data Monetization: With proper consent and privacy measures in place, there may be opportunities to monetize anonymized datasets generated through our platform. This could involve licensing data for secondary research purposes or providing insights to third-party entities in related industries.

Grants and Funding: Initially, securing grants and funding from government agencies, private foundations, and venture capital firms can provide the necessary resources to develop and commercialize our technology. As we demonstrate the value and impact of our solution, we can attract further investment to fuel growth and expansion.

By leveraging these strategies, our project has the potential to not only address critical challenges in biomedical research and healthcare but also establish a sustainable and scalable business model that delivers value to stakeholders while driving positive societal impact.

Chapter 6 Complex Engineering Problems and Activities

6.1 Complex Engineering Problems (CEP)

Table 16. Complex engineering problem attributes

Attributes		**Addressing the complex engineering problems(P) in the project**
P1	Depth of knowledge required (K3-K8)	The project requires knowledge of Designing and Simulation (K5) such as Gradio, Engineering & IT Tools (K6) such as Google Colab, Scientific Research Papers (WK8) such as Journal articles and Conference papers on Question Generation Task.
P2	Range of conflicting requirements	The more the epoch and the larger the batch size is the more computational resources such as Time and GPU memory is required.
P3	Depth of analysis required	No unique way to design. A specific framework, development environment and Interface can be selected from various alternatives such as Tensorflow/ Pytorch/Flax, Pycharm/Google Colab/Anaconda, Gradio/Flask/Docker etc.
P4	Familiarity of issues	Google Colab, PyTorch, Python, Huggingface, Gradio.
P5	Extent of applicable codes	Adapted_X work on question generation and restructured it as per our project's need and requirements.
P6	Extent of stakeholder involvement	There are several stakeholders that need to be involved including the owner of the system, Doctors, Medical Students, Educational institution, etc.

P7	Interdependence	Version compatibility of different libraries, frameworks and programming languages such as pytorch, scikit-learn, python etc.

6.2 Complex Engineering Activities (CEA)

Table 17. Complex engineering problem activities

Attributes		**Addressing the complex engineering problems(P) in the project**
A1	Range of resources	This project involves human resource, money, modern tools (Google Colab)
A2	Level of interactions	Involves interactions between different stakeholders including group members to design the system, University Students to collect data, etc.
A3	Innovation	Employs innovative skills of engineering by introducing technology in a different manner in the medical sector.
A4	Consequences to society	Impact in our society since it can be used in medical institutions as Question Answering systems are able to create questions and their exact answer efficiently for any documents.
A5	Familiarity	Needs to be familiar with a cloud-based computational environment (Google Colab).

Chapter 7 Conclusions

7.1 Summary

In this research, we developed and refined an advanced Biomedical Extractive Question Answering (BEQA) system by leveraging the strengths of transformer-based models and a suite of deep learning techniques. Our approach began by fine-tuning six pre-trained models—BERT, BioBERT, PubMedBERT, ClinicalBERT, Bio+ClinicalBERT, and BioLinkBERT—on the BioASQ dataset, with Bio+ClinicalBERT emerging as the most robust and high-performing model.

To further enhance its capabilities, we implemented a layered strategy that included transfer learning, BiLSTM integration, and ensemble learning techniques. The transfer learning phase involved fine-tuning Bio+ClinicalBERT with Named Entity Recognition (NER) and Relation Extraction (RE) tasks, allowing the model to better understand biomedical terminology and relationships within the text. The sequential fine-tuning of NER followed by RE led to the creation of the Bio+ClinicalBERT_NER_RE QA Finetuned model, which outperformed individual transfer learning approaches.

Building on this, we integrated a Bidirectional Long Short-Term Memory (BiLSTM) layer into the model to capture complex contextual dependencies, resulting in the Bio+ClinicalBERT_NER_RE_BiLSTM QA Finetuned model. This version demonstrated improved accuracy in extracting factoid answers from biomedical texts. Finally, we applied ensemble learning through a bagging approach with majority voting, producing our top-performing model, Bio+ClinicalBERT_NER_RE_BiLSTM QA Finetuned, which achieved state-of-the-art results on multiple evaluation metrics, including F1 score, exact match (EM), and lexical accuracy (Lacc).

The study underscores the importance of combining advanced techniques such as transfer learning, BiLSTM, and ensemble learning to address the unique challenges of BEQA, particularly in handling domain-specific terminology and complex scientific concepts. Our work contributes to the growing field of BEQA by providing a more accurate and reliable model that can significantly aid researchers and clinicians in extracting precise information from biomedical literature. Thus, we believe that our evaluation will help future research to investigate the integration of additional contextual and semantic features to further improve the accuracy and reliability of BEQA systems.

7.2 Limitations

The integration of advanced deep learning techniques—transfer learning, BiLSTM, and ensemble learning—has greatly enhanced BEQA models' performance. The Bio+ClinicalBERT_NER_RE_BiLSTM QA Finetuned model, in particular, demonstrated superior accuracy and robustness across all evaluation metrics. This success highlights the effectiveness of leveraging multiple models and sophisticated training methodologies to overcome the difficulties in answering biomedical extractive questions.

Despite the promising results, several limitations were identified in this study. They are:

- **Dataset Constraints**: The reliance on the BioASQ 4b and 9b datasets limited the diversity of training data. These datasets, while comprehensive, may not cover the full spectrum of biomedical knowledge, potentially limiting the model's generalizability to other biomedical domains.
- **Computational Resources**: The extensive computational resources required for training multiple models and implementing ensemble learning posed significant challenges. This constraint may limit the scalability of the proposed approach in real-world applications, particularly in resource-limited settings.

Our findings underscore the importance of task-specific enhancements and ensemble strategies in achieving state-of-the-art performance in complex domains like biomedicine. The BEQA system's ability to handle intricate queries and deliver precise answers makes it a valuable tool for researchers and clinicians who require quick access to critical biomedical information.

7.3 Future Improvement

In the future, we will extend our work to take steps including creating a relevancy classifier, conducting more extensive testing, and optimizing the system for deployment. We will also explore the ethical, legal, and social implications (e.g., privacy concerns) of utilizing deep learning techniques in the biomedical domain [38]. Moreover, we will extend our work to focus on expanding the training data, exploring other advanced architectures using Large Language Model

(LLM) [39] alongside with Retrieval Augmented Generation (RAG) [40], and testing the model on diverse biomedical QA datasets.

References

[1] M. Aggarwal, "Information retrieval and question answering nlp approach: an artificial intelligence application," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 1, no. NCAI2011, 2011.

[2] O. Kolomiyets and M.-F. Moens, "A survey on question answering technology from an information retrieval perspective," *Inf Sci (N Y)*, vol. 181, no. 24, pp. 5412–5434, 2011.

[3] Y. Sharma and S. Gupta, "Deep learning approaches for question answering system," *Procedia Comput Sci*, vol. 132, pp. 785–794, 2018.

[4] M. Lee *et al.*, "Beyond information retrieval—medical question answering," in *AMIA annual symposium proceedings*, 2006, p. 469.

[5] P. Xu, D. Liang, Z. Huang, and B. Xiang, "Attention-guided generative models for extractive question answering," *arXiv preprint arXiv:2110.06393*, 2021.

[6] M. Fajcik, J. Jon, and P. Smrz, "Rethinking the objectives of extractive question answering," *arXiv preprint arXiv:2008.12804*, 2020.

[7] A. Arbaaeen and A. Shah, "Natural language processing based question answering techniques: A survey," in *2020 IEEE 7th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, 2020, pp. 1–8.

[8] E. Sulem, J. Hay, and D. Roth, "Do we know what we don't know? studying unanswerable questions beyond SQuAD 2.0," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 4543–4548.

[9] A. Krithara, A. Nentidis, K. Bougiatiotis, and G. Paliouras, "BioASQ-QA: A manually curated corpus for Biomedical Question Answering," *Sci Data*, vol. 10, no. 1, p. 170, 2023.

[10] Q. Jin *et al.*, "Biomedical question answering: a survey of approaches and challenges," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–36, 2022.

[11] T. Stampolidou, "Extracting Local Features to Improve Transformer-based Biomedical Question Answering Models," Αριστοτλιο Πανπιστημιο Θσσαλονικη, 2022.

[12] Y. Wu, H.-F. Ting, T.-W. Lam, and R. Luo, "BioNumQA-BERT: answering biomedical questions using numerical facts with a deep language representation model," in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2021, pp. 1–6.

[13] V. Sharma, N. Kulkarni, S. Pranavi, G. Bayomi, E. Nyberg, and T. Mitamura, "BioAMA: towards an end to end biomedical question answering system," in *Proceedings of the BioNLP 2018 workshop*, 2018, pp. 109–117.

[14] M. Sarrouti and S. O. El Alaoui, "SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions," *Artif Intell Med*, vol. 102, p. 101767, 2020.

[15] S. Lee, H. Kim, and J. Kang, "LIQUID: A framework for list question answering dataset generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 13014–13024.

[16] A. Gonzalez-Agirre, M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, and M. Krallinger, "Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track," in *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, 2019, pp. 1–10.

[17] H. Wei *et al.*, "Named entity recognition from biomedical texts using a fusion attention-based BiLSTM-CRF," *IEEE Access*, vol. 7, pp. 73627–73636, 2019.

[18] Z. Zhang and A. L. P. Chen, "Biomedical named entity recognition with the combined feature attention and fully-shared multi-task learning," *BMC Bioinformatics*, vol. 23, no. 1, p. 458, 2022.

[19] D. Sousa, "Deep learning system for biomedical relation extraction combining external sources of knowledge," in *European Conference on Information Retrieval*, 2021, pp. 688–693.

[20] D. Sousa and F. M. Couto, "Biomedical relation extraction with knowledge graph-based recommendations," *IEEE J Biomed Health Inform*, vol. 26, no. 8, pp. 4207–4217, 2022.

[21] A. Lamurias and F. M. Couto, “Lasigebiotm at MEDIQA 2019: biomedical question answering using bidirectional transformers and named entity recognition,” in *Proceedings of the 18th BioNLP workshop and shared task*, 2019, pp. 523–527.

[22] S. Alrowili and K. Vijay-Shanker, “Exploring Biomedical Question Answering with BioM-Transformers At BioASQ10B challenge: Findings and Techniques.,” in *CLEF (Working Notes)*, 2022, pp. 222–234.

[23] Z. Wang and H. Guan, “Research on named entity recognition of doctor-patient question answering community based on bilstm-crf model,” in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020, pp. 1641–1644.

[24] H. Yang, M. Li, Y. Xiao, H. Zhou, R. Zhang, and Q. Fang, “One llm is not enough: Harnessing the power of ensemble learning for medical question answering. medRxiv,” 2023.

[25] X. Zhu, Y. Chen, Y. Gu, and Z. Xiao, “SentiMedQAer: a transfer learning-based sentiment-aware model for biomedical question answering,” *Front Neurorobot*, vol. 16, p. 773329, 2022.

[26] H. Kim, H. Hwang, C. Lee, M. Seo, W. Yoon, and J. Kang, “Exploring Approaches to Answer Biomedical Questions: From Pre-processing to GPT-4.,” in *CLEF (Working Notes)*, 2023, pp. 132–144.

[27] K. Peng, C. Yin, W. Rong, C. Lin, D. Zhou, and Z. Xiong, “Named entity aware transfer learning for biomedical factoid question answering,” *IEEE/ACM Trans Comput Biol Bioinform*, vol. 19, no. 4, pp. 2365–2376, 2021.

[28] A. Aksenova, T. Asamov, P. Ivanov, and S. Boytcheva, “Improving Biomedical Question Answering with Sentence-based Ranking at BioASQ-11b.,” in *CLEF (Working Notes)*, 2023, pp. 27–36.

[29] K. Yamada, M. Miwa, and Y. Sasaki, “Biomedical Relation Extraction with Entity Type Markers and Relation-specific Question Answering,” in *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, 2023, pp. 377–384.

[30] G. Tsatsaronis *et al.*, “An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition,” *BMC Bioinformatics*, vol. 16, pp. 1–28, 2015.

[31] J. D. M.-W. C. Kenton and L. K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of naacL-HLT*, 2019, p. 2.

[32] J. Lee *et al.*, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[33] E. Alsentzer *et al.*, “Publicly available clinical BERT embeddings,” *arXiv preprint arXiv:1904.03323*, 2019.

[34] Y. Ling, “Bio+ Clinical BERT, BERT Base, and CNN performance comparison for predicting drug-review satisfaction,” *arXiv preprint arXiv:2308.03782*, 2023.

[35] Y. Gu *et al.*, “Domain-specific language model pretraining for biomedical natural language processing,” *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[36] M. Yasunaga, J. Leskovec, and P. Liang, “Linkbert: Pretraining language models with document links,” *arXiv preprint arXiv:2203.15827*, 2022.

[37] W. Yoon, J. Lee, D. Kim, M. Jeong, and J. Kang, “Pre-trained language model for biomedical question answering,” in *Joint European conference on machine learning and knowledge discovery in databases*, 2019, pp. 727–740.

[38] Y. S. J. Aquino, P. Shih, and R. Bosward, “The ethical, legal and social implications of Artificial Intelligence in Public Health,” 2023.

[39] C. Anaya, M. Fernandes, and F. M. Couto, “LLM Fine-Tuning With Biomedical Open-Source Data,” 2024.

[40] C. Wang *et al.*, “BioRAG: A RAG-LLM Framework for Biological Question Reasoning,” *arXiv preprint arXiv:2408.01107*, 2024.

.